Confronting Core Issues: A Critical Test of Attitude Polarization

Yamil Ricardo Velez *Columbia University*
Patrick Liu *Columbia University*

A long-standing debate in the political psychology literature considers whether individuals update their beliefs and attitudes in the direction of evidence or grow more confident in their convictions when confronted with counter-attitudinal arguments and information. Though recent studies have shown that instances of the latter tendency, which scholars have termed attitude polarization and "belief backfire," are rarely observed in settings involving hot-button issues or viral misinformation, we know surprisingly little about how participants respond to information directly targeting deeply held attitudes, a key condition for triggering attitude polarization according to theories of motivated reasoning. To address this gap, we develop a tailored experimental design that measures participants' positions regarding their most important issues and randomly assigns them to different mixtures of personalized pro-attitudinal and counter-attitudinal information using the large language model GPT-3. We fail to recover evidence consistent with attitude polarization across three studies. We conclude by discussing implications for the study of political cognition and the measurement of attitudes.

Word Count: 9,804

Yamil Ricardo Velez is an Assistant Professor of Political Science at Columbia University, 420 W. 118th Street, New York, NY, 10027 (yrv2004@columbia.edu).

Patrick Liu is a Ph.D. student at Columbia University, 420 W. 118th Street, New York, NY, 10027 (ppl2115@columbia.edu).

Whether people process political information in an even-handed fashion or actively resist ideologically inconvenient claims is crucial for understanding citizen competence (Kuklinski et al. 2000) and democratic accountability (Achen and Bartels 2017, Chapter 10). Theories of the latter commonly assume that citizens accurately gauge economic and political conditions and incorporate these evaluations into political decisions. Deviations from these assumptions might undermine accountability if citizens attribute responsibility to politicians for events outside their control (Achen and Bartels 2017; Healy et al. 2010) or process information in service of partisan and ideological goals (e.g., Kraft et al. 2015; Jerit and Barabas 2012; Shapiro and Bloch-Elkon 2008; Gaines et al. 2007; Bartels 2002).

In their seminal article, Taber and Lodge (2006) observed troubling evidence that individuals exposed to a balanced set of pro and con arguments did not moderate their issue positions, but instead developed stronger attitudes. Nyhan and Reifler (2010) found misinformation corrections could "backfire," resulting in higher levels of factual inaccuracy relative to control. Numerous other studies support the notion that individuals assimilate congenial information and reject counter-attitudinal information. Their findings seem particularly alarming amid a contemporary media age characterized by widespread misinformation. If people double down on false notions when challenged, then extreme caution is warranted when designing corrections or encouraging political deliberation.

The ubiquity of these processes has been called into question by recent studies showing that "backfire" and "attitude polarization" are rare. Wood and Porter (2019) examine over 50 claims and find robust evidence that corrections improve belief accuracy across statements varying in partisan valence. Meta-analyses of experimental studies testing the impact of fact-checking have also revealed that beliefs generally respond to factual corrections by moving in the direction of evidence (Walter et al. 2020). Assessing both attitudes and beliefs, Guess and Coppock (2020) find no evidence of attitude polarization for politically charged topics such as gun control, minimum wage, and capital punishment.

A central priority in this more recent strain of research has been to employ politically salient issues or viral pieces of misinformation when evaluating attitude polarization and "belief backfire." But although these design choices increase the personal stakes for participants to update their beliefs, the failure in these studies to detect attitude and belief polarization may nonetheless reflect the possibility that issue areas or political claims used in extant research are not sufficiently important or accessible. Defending attitudes in the way envisioned by motivated reasoning scholars might occur only when attitudes are deeply held, stable, and personally relevant—a defense frequently invoked by past proponents (e.g., Nisbett and Ross 1980, 180). Though positive belief updating and attitude change might occur for more peripheral issues, motivated reasoning may be more likely to occur when people feel personally invested in their beliefs.

Additional design choices may have further muted the appearance of motivated reasoning in recent scholarship and, in so doing, magnified uncertainty about the state of the literature. Studies priming directional motives by imploring participants to take partisan goals into account (Bayes et al. 2020) or encouraging the activation of "on-line processing" – as opposed to "memory-based processing" – have found evidence for motivated reasoning (Redlawsk 2002). Studies of attitude polarization have also had to confront measurement challenges, as ceiling effects may obscure the presence of attitude polarization when respondents score highly on measures of attitude strength.

We conduct a critical test of the attitude polarization hypothesis that addresses these many critiques. Most notably, our design targets personal issue importance by (i) measuring participants' most deeply held issue attitudes and (ii) exposing them to personally relevant counter-attitudinal information and mixtures of pro- and counter-attitudinal information. We accomplish this by leveraging a state-of-the-art large language model (LLM), GPT-3, that is capable of constructing personally tailored attitude measures and political arguments on the fly. In addition, we (iii) encourage participants to engage in directional motivated reasoning and (iv) are careful to construct outcome measures that reduce ceiling effects. Across two studies priming directional motives and one study varying issue strength, we fail to find evidence of attitude polarization despite observing striking evidence of intervening mechanisms consistent with the theory of motivated reasoning. We also observe robust evidence that our tailored approach captures strong attitudes that are confidently held and stable over time. We conduct a third experiment that exposes participants to longer and more affectively charged counter-arguments, and detect evidence of substantively large *decreases* in attitude strength (.26 control group standard deviation units). Our study sets up ideal conditions for detecting attitude polarization and fails to recover an effect. We conclude by discussing how the use of tailored interventions and outcome measures can improve our understanding of political behavior and persuasion.

Motivated Reasoning and Related Processes

Evidence of motivated reasoning as widely pervasive has been elusive under the recent empirical literature, as studies suggest that attitude polarization is “the exception, not the rule” in political learning (Guess and Coppock 2020, 1500). Below, we recount how the litmus test for detecting polarization has grown increasingly difficult to satisfy as detailed studies have failed to replicate the phenomenon. We pinpoint key assumptions in the theoretical foundations of the attitude polarization hypothesis deserving deeper inquiry.

Theories of “selective perception,” which include motivated reasoning, share the view that prior convictions heavily influence people’s interpretations of information and events but vary in the degree to which they ascribe selective perception to motivational or cognitive mechanisms. Festinger’s (1957) cognitive dissonance theory underscored how the costs of holding inconsistent attitudes motivate resistance to contrary information. Nisbett and Ross (1980) took issue with this portrayal, arguing for a nonmotivational view of inferential errors as merely inevitable in human cognition.

Motivated reasoning is primarily associated with the formulation of Kunda (1987, 1990), under which an individual confronted with new information is motivated to reach either an accurate conclusion when pursuing *accuracy goals* or a specific conclusion—typically one that minimizes friction with prior beliefs—when pursuing *directional goals*. These goals determine how individuals perceive, process, and respond to political information inconvenient to one’s worldview. Taber and Lodge (2006) brought motivated reasoning further into the mainstream and drew special attention to the role of affect. They pointed to ”hot cognition,” the hypothesis that previously evaluated sociopolitical concepts remain affectively charged in memory (Redlawsk 2002; Lodge and Taber 2005). Reexposure to a concept automatically and instantaneously activates the accompanying affect, which in turn generates directional motivations.

Taber and Lodge (2006) crystallized motivated reasoning’s constitutive processes into three hypotheses: prior attitude effect, disconfirmation bias, and confirmation bias. The first two are particularly relevant to attitude polarization. Disconfirmation bias predicts an inclination to counterargue and denigrate attitudinally incongruent arguments more than congruent arguments. The prior attitude effect, whereby those who hold core issue preferences evaluate ideologically congruent arguments as stronger than incongruent ones, finds roots in Lord et al.'s (1979) classic study. There, students who supported and opposed the death penalty were presented details about two studies on its deterrent effects, one touting its efficacy and the other undermining it. Participants regarded the study that affirmed their prior view as significantly better designed and more convincing than the study opposing their view.

Detecting Attitude Polarization

The central prediction of motivated reasoning is that when directional motivations are triggered, people will preserve or even strengthen their prior attitudes to defend against counter-attitudinal evidence. This concept of attitude polarization drew heightened attention with Lord et al.'s (1979) demonstration, in which students self-reported increases in their opinion strength after being exposed to mixed evidence for the deterrent effects of capital punishment. Lord and colleagues attributed this finding to *biased assimilation*, wherein individuals accept evidence that supports their initial view but discount evidence that contradicts. At scale, biased assimilation entails the divergence of opinions between people who hold opposing views when presented the same mixture of pro- and counter-attitudinal information.

Though many scholars criticized the use of self-described opinion change by Lord et al. (1979) and failed to replicate polarization when they measured actual pretreatment and posttreatment opinions (e.g., Miller et al. 1993; Kuhn and Lao 1996; Munro and Ditto 1997), Taber and Lodge (2006, 756) argued that these studies may simply have "failed to arouse sufficient partisan motivation to induce much biased processing." After displaying impassioned pro and con arguments by real interest groups to Stony Brook students with strong prior attitudes about affirmative action and gun control, Taber and Lodge found that many students reported higher levels of post-treatment attitude strength. They characterized these results as most consistent with the affect-based explanations offered by motivated reasoning. When Nyhan and Reifler (2010) found that corrections to misinformation strengthened the convictions of certain strongly committed subjects – a result they coined the "backfire effect" – they likewise relied on theories of affective and motivated reasoning (323), speculating that vigorously counterarguing belief-incongruent information can bring to the front of one's mind more congenial considerations that bolster prior beliefs. A wide range of studies have observed polarization and backfire on varied issue topics, often among narrow subgroups (e.g., Hart and Nisbet 2012; Nyhan and Reifler 2015; Zhou 2016; Schaffner and Roche 2017; Ecker and Ang 2019).

The most recent available evidence, however, raises real doubt about whether polarization is but a rare occurrence. Wood and Porter (2019) failed to replicate backfire effects despite testing 52 issues across 5 experiments with more than 10 thousand subjects, as did Guess and Coppock (2020) across three large survey experiments. Haglin's (2017) direct replication of Nyhan and Reifler (2015) likewise failed to detect backfire. Aggregating conclusions from these and numerous other studies that have failed to replicate the backfire phenomenon (e.g., Garrett et al. 2013; Weeks 2015; Nyhan et al. 2020), Swire-Thompson et al.'s (2020, 288) review offers three potential explanations for its elusory nature: "either (a) the backfire effect is difficult to elicit on the larger group level, (b) it is extremely item-, situation-, or individual-specific, or (c) the phenomenon does not exist at all." It remains a subject of ongoing debate whether, in theory, Bayesian updating can be adjusted to accommodate instances of attitude polarization and under what conditions polarization can offer distinctive evidence of motivated reasoning (e.g., Benoît and Dubra 2019; Little 2021).

Revisiting Attitude Strength in Motivated Reasoning

In this paper, we seek to advance the debate on attitude polarization by drawing renewed attention to the role of attitude strength in nurturing the conditions that purportedly give rise to motivated reasoning. That subjects experience a deep-seeded desire to protect attitudes that are affect-laden or to which they assign great personal importance has long been presumed essential to the phenomenon, if not necessary for its occurrence. Taber and Lodge (2006, 757) termed this the *attitude strength effect*, whereby "those citizens voicing the strongest policy attitudes will be most prone to motivated skepticism."

At a high level, psychologists have defined four features of strong attitudes: resistance to change, stability over time, impact on judgment and cognition, and impact on behavior (Krosnick and Petty 1995). Different accounts have varied substantially how they frame the relationship between strength and other features of attitudes. As Boninger et al. (1995) note, psychologists long considered strength more a metaphor than a formal construct and measured attitude extremity, accessibility, certainty, and commitment each as indicators for strength. In a more recent review of the subject, Howe and Krosnick (2017) describe these many terms as standalone concepts that cooccur with strength in complex ways, while nonetheless subsuming them under the phrase "strength-related features."

For our purposes, we will speak broadly about attitude strength as an umbrella concept, relying where appropriate on such related factors as importance and certainty. Our reasoning is twofold. First, a wide-ranging literature largely predating Taber and Lodge (2006) explored potential relationships between various strength-related features of attitudes and the constituent processes and outcomes of motivated reasoning. According to these studies, attitude polarization may be exacerbated or more likely to occur among those who assign their relevant attitude greater levels of importance (Tesser and Leone 1977), commitment (Pomerantz et al. 1995), and extremity (McHoskey 1995), while the prior attitude effect may be particularly pronounced among those whose relevant attitudes are personally important (Clore and Baldridge 1968; Krosnick 1988, 1990) and highly accessible (Houston and Fazio 1989). While nuanced differences abound, the literature suggests substantial overlap between these strength-related features, particularly in the ways they bear on motivated reasoning. A second rationale relates to the empirical strategy in this paper. The theory of motivated reasoning as elaborated by Taber and Lodge (2006) relies on attitude strength without disambiguating between these varied features. We imitate their approach and consider a range of strength-related concepts and measures in the interest of constructing a generous test of the attitude polarization hypothesis.

Individuals may assign personal importance to issues and subsequently develop strong attitudes about them for heterogeneous reasons. The degree to which an individual perceives an issue as bearing upon their self-interest, social identification with important reference groups (i.e., partisan identification), and cherished social and personal values determines the level of personal importance allotted to that issue (Boninger et al. 1995). Highly important attitudes, in turn, might generate the motivations that induce selective information processing (Lavine et al. 2000). There is suggestive evidence that more important attitudes are more resistant to change (Gopinath and Nyer 2009; Zuwerink Jacks and Devine 1996), and corrections to misinformed beliefs may be less effective when those beliefs are perceived as personally important (Vidigal and Jerit 2022).

The key question taken up in this paper is whether strong attitudes and weak attitudes operate so differently on reasoning that motivated reasoning might only be observed when the former are at play. That is, would an experimental design that elicits subjects' most deeply held attitudes unearth evidence of attitude polarization? Taber and Lodge (2006, 754) questioned whether earlier studies had dismissed attitude polarization primarily because their selection of arguments and evidence were insufficiently affect-laden and thus unable to arouse the requisite motivations. When Nyhan and Reifler (2010, Study 2) corrected misperceptions about the discovery of weapons of mass destruction during the U.S.'s 2003 invasion of Iraq, their post hoc analysis found backfire effects present only among a subset of conservative respondents who rated Iraq as the most important issue for the U.S., leading the authors to similarly conclude that backfire effects are contingent on issue importance.

Several recent papers have endeavored to resolve this lingering uncertainty, whether by testing a great assortment of issues of ”keen political interest” and with partisan valence (Wood and Porter 2019, 142) or by selecting issues made salient by the latest national news. For instance, Guess and Coppock (2020, Experiment 1) administered one study of backfire effects using information about gun safety just days after the 2016 Orlando mass shooting. Those studies found little evidence of polarization in spite of research designs generous toward the motivated reasoning hypothesis.

It is not clear, however, that personal salience and national salience can be merged as concepts. As Ryan and Ehlinger (2019, 4) argue, there is no consensus to date about what number of political topics the typical person cares about, nor the number of distinct topics that matter to the electorate at large. Not every issue of national import will induce in every individual the need to put up defenses against discordant evidence. A critical test of attitude polarization instead necessitates a research design that can, first, invite open-ended input from participants about the issues they hold deeply important, and second, assess the efficacy of persuasion on these core attitudes by confronting participants with tailored pro- and counter-attitudinal responses.

Tailoring Information with GPT-3

We perform such a critical test by taking advantage of advances in large language models—dense neural networks with high levels of performance in replicating human speech. Though training language models is time- and resource-intensive, *pre-trained*, *task-agnostic* models with general applicability and that can be adapted to specific NLP tasks with far fewer resources are growing increasingly accessible, presenting a window of opportunity for experimental social science. GPT-3, an autoregressive language model with 175 billion parameters developed by the AI research lab OpenAI, is one such model. When fed a natural language instruction and perhaps a small number of demonstrations ("in-context learning"), GPT-3 can show strong performance both in on-the-fly reasoning and in text completion tasks such as translation and question-answering (Brown et al. 2020). As Brown and colleagues illustrate, in the most minimal form of in-context learning—what they term "zero-shot learning"—GPT-3 can deliver when provided only a brief task description.

Even as pre-trained language models face enduring technical limitations, recent advances evince their long-term potential for political science methodology. Porter and Velez (2022) demonstrated how using GPT-2 (GPT-3's predecessor) to automate a "placebo sampling" process aids in measuring survey experimental treatment effects. Here, we illustrate two novel applications of pre-trained language models. First, using participants' open-ended responses detailing issues of personal importance, GPT-3 was able to construct tailored 7-point Likert items measuring participants' attitude strength and certainty about those issues. Second, we demonstrate GPT-3's capacity to generate a suite of persuasive arguments when provided only (a) a topic of political discussion by the research participant and (b) brief, issue-agnostic instructions from the researcher about the position to be taken (pro/con). The ability to develop personalized measurement scales and tailor persuasive messages in the particular context of large online experiments advances the persuasion literature beyond extant strategies for confronting personal issue importance.

Constructing a Critical Test

In addition to eliciting important issues and tailoring persuasive treatments, our experiments involve several other design choices aimed at raising the chances of detecting polarization and determining its origins. Failing to detect attitude polarization despite all these conditions provides strong evidence that the phenomenon may occur only rarely. We briefly summarize these design choices.

1. **Strong attitudes:** Respondents provide open-ended input about their deeply held attitudes. We validate the strength of these attitudes using multiple measures.

2. **Tailored arguments:** We employ GPT-3 to generate arguments tailored to respondents' strong attitudes. We test both single-sentence arguments (Experiments 1 and 2) and full paragraphs (Experiment 3).

3. **Motivational primes:** We randomly assign participants to be primed either for directionally motivated reasoning or for accuracy motivations prior to argument exposure in Experiments 1 and 2.

4. **Multiple information conditions:** We randomly assign participants to view and respond to four pro-attitudinal arguments, four counter-attitudinal arguments, or two of each. Past studies have typically included only a Con or a Mixed condition. The Mixed argument condition offers a potential test of the biased assimilation hypothesis à la Lord et al. (1979), while evidence of polarization under Mixed or Con conditions would align with the motivated reasoning view.

5. **Multiple outcome measures:** Studies of attitude polarization often face limitations due to ceiling effects. Similar to Taber and Lodge (2006), we minimize this issue by employing multiple outcome measures as well as a multi-item scale of attitude certainty, improving our chances of detecting the strengthening of already strong attitudes.

6. **Multi-wave design:** Consistency effects might mute the presence of polarization when attitudes are measured before and after the treatment. We mitigate such concerns in Experiment 2 by using a two-wave design. Experiment 3 uses a single-wave design where attitudes are only measured after the intervention.

7. **Measuring intervening processes:** We validate whether our experimental design successfully set conditions for triggering motivated reasoning using measures of two theorized mechanisms of motivated reasoning—prior attitude effect and disconfirmation bias.

Data, Methods, and Results

We develop a tailored experimental design that constructs personalized interventions and outcome measures using open-ended responses. The basic structure of the experimental design is the following: (1) participants report their position on a core political issue using an open-ended question, (2) text from this question is passed to GPT-3, returning a one-sentence summary and eight arguments related to the issue position, (3) participants report attitude strength and certainty with respect to this deeply-held issue position, (4) participants are randomly assigned to different combinations of pro- and counter-attitudinal information relevant to their issue position after being primed to either consider the arguments from a directional or accuracy-minded perspective, and (5) attitude strength and certainty are reassessed. Experiment 1 implemented this approach in one wave using a pre-post design (Clifford et al. 2021). Experiment 2, carried out in two waves, assessed attitudes toward core and peripheral issues to evaluate variation in effects across levels of attitude strength. A modified design was employed in Experiment 3 to explore how the core issue attitudes we elicit are influenced by more heavy-handed persuasive messages.[1]

In Appendix B.8, we show that across the three experiments, participants' self-reported core attitudes map onto 66 unique issue topics. Although abortion and other salient issues are mentioned frequently, no single issue accounts for more than a quarter of responses. Eliciting open-ended responses avoids assuming any particular topic of national salience will elicit personal salience in our participants.

Experiment 1

We recruited 2,141 participants using the online survey platform CloudResearch Connect (CR).[2] The survey was in the field from September 28, 2022 to October 1, 2022.[3] Upon reading the consent

[1] All of the experiments were reviewed and approved by the institutional review board at the authors' institution (Protocol #AAAU3638).

[2] To ensure high-quality open-ended responses and compliance with the thought-listing, we sought a survey vendor with especially attentive subjects. Online convenience samples often replicate ATEs and CATEs observed using nationally representative samples (Mullinix et al. 2015; Coppock et al. 2018).

[3] The pre-analysis plan for Experiment 1 is available here: https://aspredicted.org/3Z5_F2C

form and agreeing to participate, participants were taken to the following open-ended question: "Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? For example, if you care about farm subsidies, you can write 'I believe farm subsidies should be increased to help farmers.' Please write a brief sentence about an issue that you care about and where you stand on the issue." Their response was passed to OpenAI's GPT-3 text completion API, a one-sentence summary was produced, and this summary was presented as a 7-point Likert item, ranging from "strongly disaree" to "strongly agree." In 17% of cases, GPT-3 was unable to provide output. Output in all experiments was passed through OpenAI's content filter. Following OpenAI's guidelines, we prevented GPT-3 from producing "toxic content" by calling the API until the content filter did not trigger. If this condition could not be met, we flagged the observation in Qualtrics and provided a generic set of arguments. As noted in our pre-analysis plan, we exclude these cases because participants did not receive tailored information. This leaves us with an effective sample size of 1,782.[4] The total cost of using the GPT-3 API for this study was $20.49.

The text completion API was instructed to summarize each statement in one sentence. For example, one participant wrote "congress should address the issue on health care cost." The GPT-3 completion API produced the following Likert item for this respondent: "To what extent to do you agree or disagree with the following statements? I believe that Congress should address the issue of health care costs." After responding to a personalized Likert item, participants were asked about their level of certainty regarding this issue position on a 0-100 scale.

After measuring attitude strength and certainty, participants were randomized to one of two motivational conditions. These conditions described the thought-listing task, but emphasized engaging with the argument from a more fair-minded perspective ("accuracy" motivation) or from the perspective of maintaining consistency ("directional" motivation). Those in the former condition were explicitly instructed to "ignore any personal feelings or emotions" and focus on the "truth of each

[4] In Appendix B.4, we analyze whether certain issues are less likely to receive valid GPT-3 output. First, we note that the API calls occur before treatment, thus avoiding confounding. Second, though there is modest variation across issues with respect to error-free completion rates, we fail to reject the null of equal completion rates ($F(14, 1310) = .79$, p = .68).

TABLE 1. Representative GPT-3 output

Open-ended response	congress should address the issue on health care cost.
likert	I believe that Congress should address the issue of health care costs.
pro1	Health care costs are a burden on many Americans
pro2	Reducing health care costs could free up money for other important programs
pro3	Addressing the issue of health care cost could improve the quality of life for many Americans.
pro4	Congress has a responsibility to represent the people and their interests.
con1	There is no one-size-fits-all solution to reducing healthcare costs.
con2	Some solutions to reducing healthcare costs could be unpopular with voters.
con3	Addressing the issue of healthcare cost could be costly in itself.
con4	It is not clear that Congress has the power to directly address healthcare cost.

statement," whereas those in the latter condition were instructed to "not worry" about the accuracy of each statement but instead focus on understanding what the statement means to them, given their existing beliefs. Participants were then randomized to one of three information conditions for the thought-listing task, exposing them to four pro-attitudinal arguments ("Pro" condition), four counter-attitudinal arguments ("Con"), or two pro- and two counter-attitudinal arguments ("Mixed"). Arguments were produced by passing a summary of the participant's issue position to GPT-3's text completion API. This output was returned as an eight-item JSON file that could be displayed in Qualtrics. In total, four pros and four cons were generated for each open-ended response. In the "Pro" ("Con") condition, respondents would see all of the pro (con) arguments. In the "Mixed" condition, respondents would see two randomly selected pro and con arguments. A participant who strongly supported universal healthcare, for example, could see pro-attitudinal arguments such as "it's unfair that people have to choose between basic needs and medical care" or counter-attitudinal arguments such as "there would likely be long wait times for non-urgent medical procedures if everyone had equal access to affordable healthcare," depending on the information condition. Each argument was presented with a corresponding text box below it so that participants could write their thoughts about each argument.[5] Attitude strength and certainty were then measured using

[5] The full list of arguments can be found here: bit.ly/3ZlZu26.

the same procedure as in pre-treatment. After this section of the experiment, participants filled out a brief demographic battery and read a debriefing statement revealing that the information they saw was produced by GPT-3.

Models

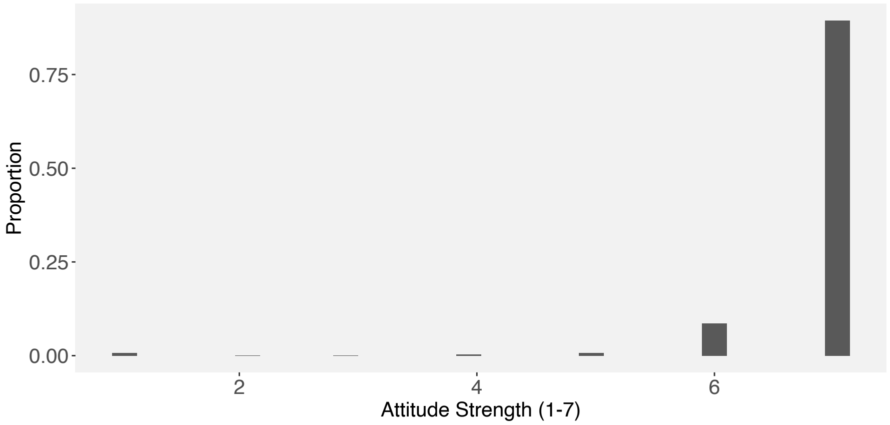
We regress post-treatment measures of attitudes (i.e., attitude strength and attitude certainty) on *pre-treatment strength and certainty measures*, information condition indicators, motivation condition indicators, and their interaction using OLS regression. Due to the inclusion of pre-treatment measures, these models capture within-study variation in attitudes. The models take the following form:

$$\begin{aligned} Y_i = \alpha + \beta_1 \text{Mixed}_i + \beta_2 \text{Con}_i + \beta_3 \text{Pre-Treatment Strength}_i + \beta_4 \text{Pre-Treatment Certainty}_i + \\ \beta_5 \text{Mixed}_i \times \text{Directional Prime}_i + \beta_6 \text{Con}_i \times \text{Directional Prime}_i + \epsilon \end{aligned} \tag{1}$$

For ease of interpretation, we plot treatment effects across the various information and motivation conditions. Based on the previous literature, our key expectation is that exposure to counter-attitudinal information in the “Con” and “Mixed” conditions should increase attitude strength and polarization relative to the “Pro” condition. Although previous research has stressed the importance of balanced information in activating motivated reasoning, we are agnostic regarding the ordering of effect sizes for the “Con” and “Mixed” conditions and consider *any* positive estimates associated with the “Con” condition (β_2, β_6) or “Mixed” condition (β_1, β_5) as evidence of attitude polarization.

Before moving on to our key analyses, we conduct validation tests of whether intervening processes such as disconfirmation bias can be detected in our experiment (see Appendix A.4). We find that those exposed to pro-attitudinal arguments generally spend approximately 5% less time on the thought-listing task than those exposed to counter-attitudinal arguments (SE = .007; p < .001). The share of denigrating responses to arguments also increases from about 9% in the “Pro” condition to 25% in the “Con” condition (SE = .019; p < .001). These results are consistent with an extensive literature suggesting that people expend more cognitive effort when considering arguments inconsistent with their prior beliefs and attitudes.

FIGURE 1. Histogram of Pre-Treatment Attitude Strength (Experiment 1)



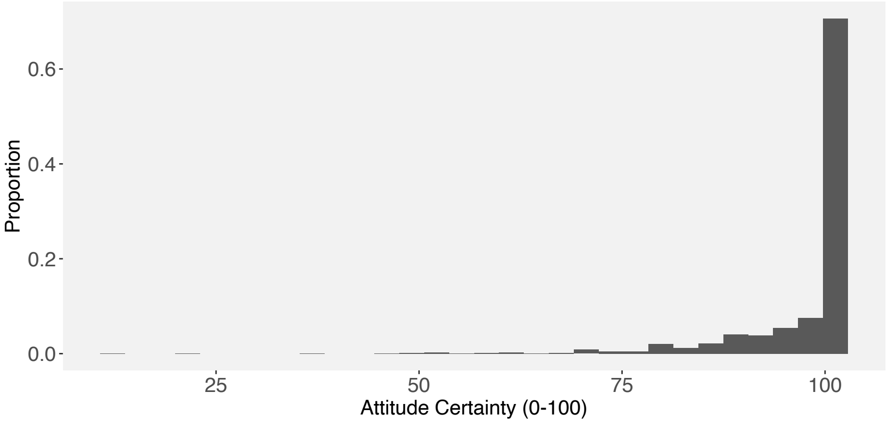
Note: Mean attitude strength is 6.83, with a standard deviation of .65.

We now consider the effects of the different experimental conditions on attitude strength. Recall, participants provided us with open-ended responses to a question asking about their most deeply-held issue position and we used GPT-3 to generate a personalized 7-point Likert item. For reference, Figure 1 displays the histogram for pre-treatment attitude strength. The mean attitude strength was 6.83, with 89% of participants selecting the maximum value of attitude strength. If attitudes are polarizing due to exposure to counter-attitudinal information, we ought to expect a larger point estimate in the "Mixed" and "Con" conditions relative to the "Pro" condition. If attitudes are moderating, estimates should move in a negative direction.

As shown in Figure 3, differences across information and motivation conditions are small, negative, and statistically indistinguishable from zero. Among participants primed to consider accuracy, those in the "Mixed" and "Con" condition score .016 (SE = .036) and .023 (SE = .039) scale points lower on attitude strength than those in the "Pro" condition, respectively. For those assigned to the directional prime, differences between the "Pro" and "Mixed" and "Con" conditions are -.022 (SE = .034) and -.027 (SE = .034) scale points, respectively. In sum, we do not find evidence that the different information and motivation conditions are shifting attitudes.

We next turn to our analyses of the 101-point attitude certainty scale. Figure 2 displays the distribution of pre-treatment attitude certainty. The mean attitude certainty score is 97. 71% of the sample selected the maximum certainty score, while 90% provided a score above 90. As shown in

FIGURE 2. Histogram of Pre-Treatment Attitude Certainty (Experiment 1)

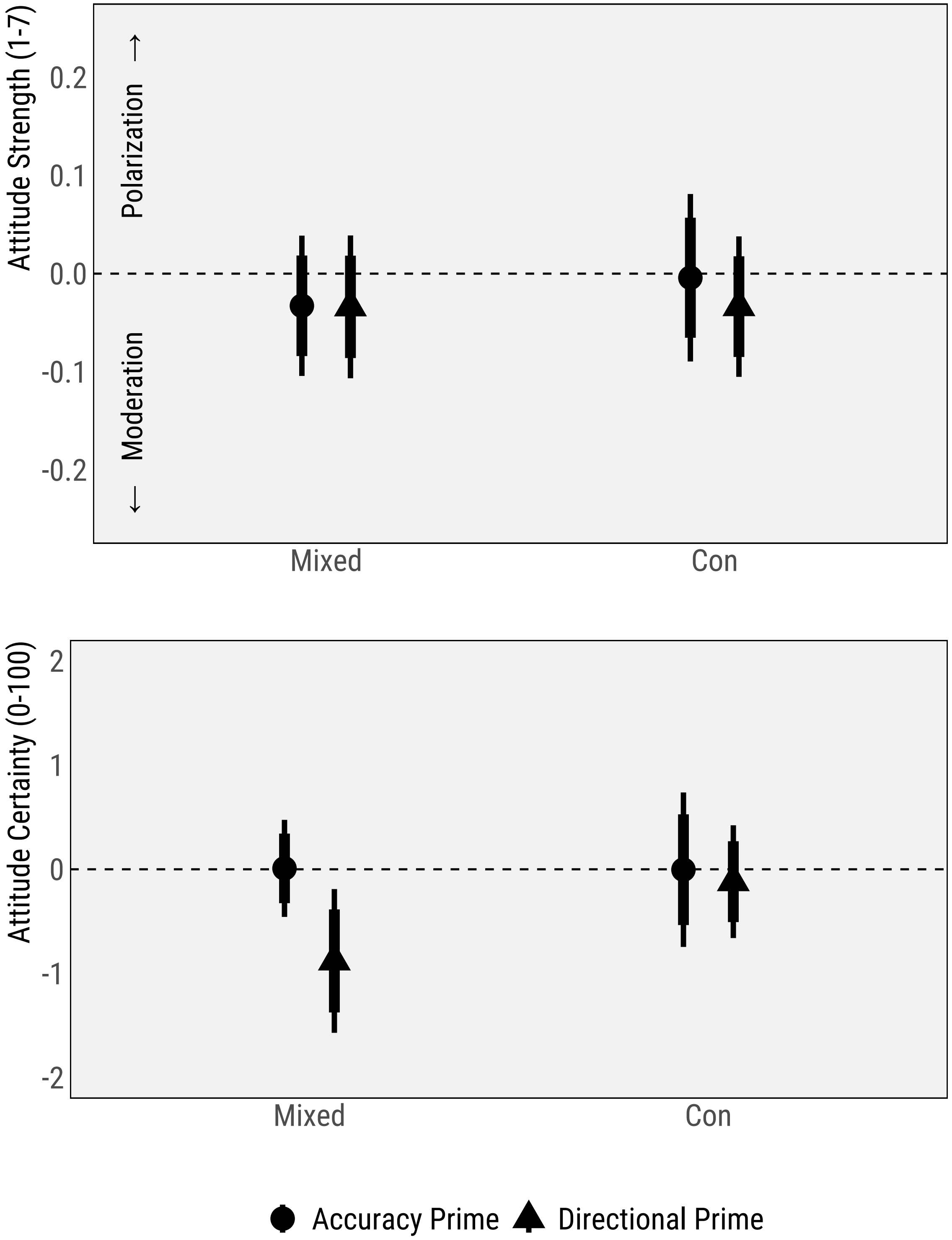


Note: Mean level of attitude certainty is 97 with a standard deviation of 7.7.

Figure 3, there are minimal differences between information and motivation conditions. For those in the accuracy condition, the difference between the "Pro" and "Con" condition is -.013 (SE = .37) scale points, and the difference between the "Pro" and "Mixed" condition is .04 (SE = .24) scale points. Those in the directional condition evince a slightly different pattern. Those in the "Mixed" condition score approximately .88 scale points lower on attitude certainty relative to those in the "Pro" condition (SE = .35; p = .01). This corresponds to a shift of .12 standard deviations, and runs in the opposite direction of what the motivated reasoning paradigm leads us to expect: those in the "Mixed" condition should express *more* certainty in their attitudes, relative to those who receive pro-attitudinal information.

In our validation tests, we detect evidence of disconfirmation bias, with those in the "Con" condition responding more critically to their assigned arguments than those in the "Pro" and "Mixed" conditions. As mentioned above, a mere 9% of those in the "Pro" condition denigrated arguments compared to 25% in the "Con" condition. Moreover, we detect evidence that participants spend 5% longer on counter-attitudinal arguments relative to pro-attitudinal arguments when completing the thought-listing task. However, despite providing arguments that target core issues and priming participants to consider information from a directional perspective, we fail to uncover evidence of attitude polarization across two measures. In one case, we observe reductions in certainty, which runs counter to the expectations of the motivated reasoning paradigm.

FIGURE 3. Effect Estimates on Attitude Strength and Certainty



Note: This figure presents point estimates and confidence intervals for attitude strength and certainty. Thick bands are 84% confidence intervals, used to facilitate visual comparisons of coefficients. Thin bands are 95% confidence intervals.

That being said, certain design choices could explain the non-findings. First, we conducted the study in a single wave, with pre- and post-intervention measures of outcomes spaced several questions apart. Although this can improve precision without significantly altering treatment effects (Clifford et al. 2021), participants may feel compelled to report a consistent outcome value within the same survey, which would hinder our ability to detect effects. Second, even in the absence of a "consistency bias," our design might foreclose the possibility of detecting polarization due to ceiling effects. We included a continuous measure of attitude certainty for this reason. However, the baseline level of certainty in our sample was 97 on a 0-100 scale. Thus, even with a finer-grained measure, we may still have trouble detecting positive treatment effects.

Experiment 2

We conducted a multi-wave experiment on CR to address the aforementioned limitations. Wave 1 was in the field from October 10, 2022 to October 13, 2022 (N = 1,591).[6] In Wave 1, we obtained pre-treatment measures of attitude strength, certainty, duration, and discussion frequency for core and peripheral issues. We measured core issue attitudes using the open-ended approach described in Experiment 1. To assess if effects differed depending on attitude strength, we also measured attitudes toward more peripheral issues. Respondents selected from a prepared list of 10 issues and responded to attitudinal questions (e.g., strength, certainty). To ensure that we were measuring weaker attitudes, we explicitly asked participants to select issues that they follow, but otherwise do not have a strong opinion on. The ten issues included foreign aid, school funding, inflation reduction, public transportation, universal healthcare, free speech, gun control, minimum wage, student loans, and marijuana legalization. Participants could choose any side of the issue.

For both core and peripheral issue positions, we measured attitude strength (7-point Likert item), attitude certainty (101-point certainty scale), and "attitude clarity and correctness" (seven items). Following Petrocelli et al. (2007), each item in the "attitude clarity and correctness" scale is a nine-point Likert item that ranges from 1 (not at all certain) to 9 (very certain).[7] This scale scores high on reliability (α = .92 for the peripheral issue; α = .96 for the core issue). After a one-week washout

[6] The pre-analysis plan for Experiment 2 can be found here: https://aspredicted.org/C4T_FG2

[7] As shorthand, we refer to this scale as "multi-item certainty."

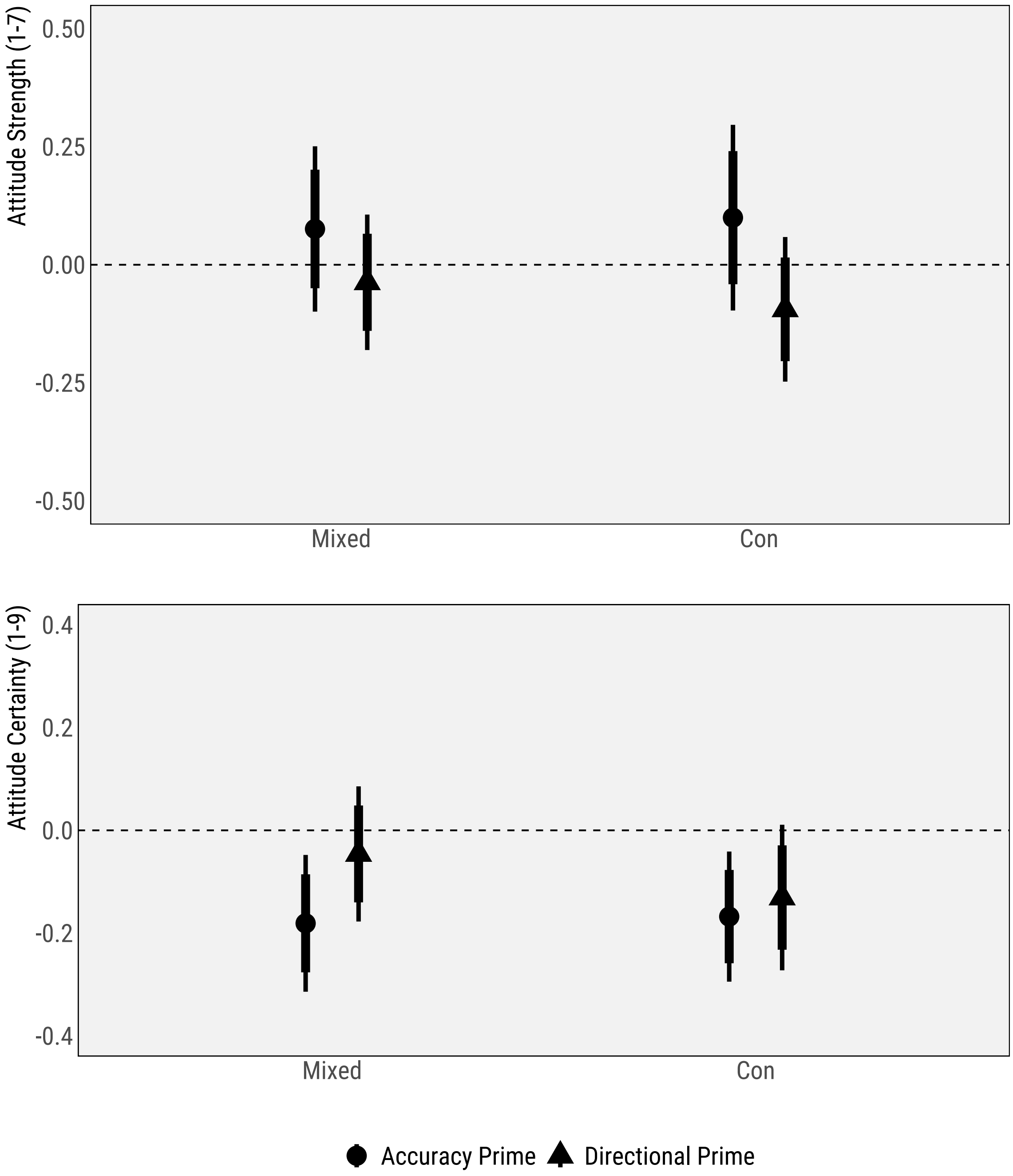
period, we carried out randomization and post-treatment outcome measurement in Wave 2 (N = 1,313).

We conducted two within-subjects trials by issue (weak vs. strong) that randomly assigned participants to one of two primes (directional vs. accuracy) and one of three information conditions (100% pro-attitudinal arguments, 100% counter-attitudinal arguments, or 50% pro- and counter-attitudinal arguments). Thus, for the first trial, a respondent could have been assigned to a peripheral issue with a directional prime and counter-attitudinal arguments, whereas for the second trial, she could have been assigned to a core issue with an accuracy prime and a mixture of pro- and counter-attitudinal arguments. We randomized the order of the issue condition (i.e. strong vs. weak). Within each trial, we measured strength and certainty after the thought-listing task, as in Experiment 1. However, in Experiment 2, we also included multi-item scales capturing argument strength and factual accuracy. These scales followed the attitude strength and certainty measures. After completing both trials, we measured demographics, participants were thanked for their participation, and everyone was debriefed. Of the 1,313 participants who completed Wave 2, 1,137 received a valid GPT-3 response for their "core issue" and 1,225 received a valid GPT-3 response for their "peripheral issue." The total API cost was $35.18. Below, we focus on the findings with respect to the "core issue" before moving on to our discussion of the "peripheral issue" findings.[8]

The top panel of Figure 4 presents effect estimates on attitude strength for core issues. We observe little attitude change across the information and motivation conditions. In four out of the four tests, effect sizes are small and statistically indistinguishable from zero. We now consider the multi-item certainty measure. As was the case with the other attitudinal measures in Experiment 1, scores are generally at the upper end of the distribution, with a mean score of 7 and a standard deviation of 2. Focusing on this outcome, we find evidence of moderation when comparing the "Con" and "Mixed" conditions to the "Pro" condition. We detect shifts of -.13 (SE = .07) and -.05 (SE = .07) scale points in the "Con" and "Mixed" condition relative to the "Pro" condition when respondents are primed to operate in a directional mode. These estimates are magnified in the accuracy condition, such that those the "Con" and "Mixed" conditions score .17 (SE = .07) and .18

[8] The full list of arguments can be found here: bit.ly/3GR0Z15

FIGURE 4. Effect Estimates on Attitude Strength and Certainty



Note: This figure presents point estimates and confidence intervals for attitude strength across information and motivation conditions with facets defined by issue attitude strength. Thick bands are 84% confidence intervals, used to facilitate visual comparisons of coefficients. Thin bands are 95% confidence intervals.

(SE = .07) scale points lower on certainty than those in the "Pro" condition. This is equivalent to a movement of approximately .09 standard deviations on the outcome. Importantly, in Appendix B.1, we report an exploratory change score analysis demonstrating that these shifts reflect those in the “Con” and “Mixed” condition moderating, rather than those in the “Pro” condition growing more confident. By and large, participants in the “Pro” condition either moderated or showed little change between the pre- and post-treatment measures of attitude strength and certainty on their core issues. In sum, we observe attitude change inconsistently, but when we do, estimates uniformly move towards moderation, instead of polarization.

In Experiment 2, we fail to detect evidence of attitude polarization. Although we observe moderate differences in our outcomes due to different information conditions, we do not detect evidence that those in the “Mixed” and “Con” conditions become more extreme or certain relative to those in the “Pro” condition. Instead, we find some evidence of moderation in the presence of counter-attitudinal arguments, especially when participants are reminded to be accurate.

Potential concerns

Across two studies that primed directional motives, varied exposure to different kinds of information, and targeted deeply-held issue positions, we failed to detect attitude polarization. Despite this, one might object that we have not successfully measured strong attitudes or that our information interventions are too weak. We address these potential concerns below.

Insufficiently strong attitudes

Might our failure to observe attitude polarization indicate we are still assessing weak attitudes? As a benchmarking exercise, we compare core and peripheral issues across several dimensions to assess if we can recover a distinct set of issue attitudes using our open-ended method. Mean responses on a 7-point Likert item are 5.03 (s = 1.29) for the peripheral issue and 6.76 (s = .78) for the core issue, respectively. Using the 0-100 certainty scale, the mean response was 96 (s = 9.15) for the core issue and 65 (s = 23) for the peripheral issue. Focusing on multi-item certainty, mean responses are 8.41 (s = .88) and 5.8 (s = 1.96). Across all scales, attitudes measured using our open-ended method possess averages close to the maximum.

76% of respondents report that they have held their core issue attitude for more than four years, while this number drops to 41% for the peripheral issue attitude (16% reported forming the peripheral attitude in the middle of the survey versus 2% for the core issue). 76% of respondents report “never” speaking about the peripheral issue in the past six months, compared to 37% for the core issue.

Examining attitude stability, approximately 44% of respondents retain the same level of attitude strength across waves for the peripheral issue, while this number is 81% for the core issue. When subsetting on those who report the maximum level of attitude strength in Wave 1, 62% of respondents remain at the maximum for the peripheral issue, compared to 87% for the core issue. In sum, our evidence suggests attitudes pertaining to the core issue are stable, durable, and personally relevant.

We detect strong evidence of a prior attitude effect. As we report in Appendix Section A.6, we find that people generally rate pro-attitudinal (counter-attitudinal) arguments as stronger (weaker) when they possess stronger attitudes toward a given topic. We also find that the gap between pro and con ratings is larger for core versus more peripheral issues. In Appendix Section A.7, we also successfully replicate disconfirmation bias, whereby individuals expend cognitive resources on combatting counter-attitudinal information. We find that participants generally spend more time on the thought listing and are more likely to denigrate arguments when they challenge (versus support) one's pre-existing attitudes.

In sum, our results indicate that we have measured attitudes that are strong enough to provoke an attitudinal defense. Across different descriptive statistics, we find that the tailored measurement approach generally yields responses at the upper end of stability, strength, certainty, duration, and discussion frequency. We also uncover evidence of a prior attitude effect, where respondents rate counter-attitudinal arguments as weaker and less accurate when those arguments target a core versus peripheral issue. Finally, we detect differences between strong and weak attitudes with respect to disconfirmation bias, with participants devoting more time to counter-attitudinal arguments and expressing more disagreement when arguments target a core versus peripheral issue position.

Insufficiently strong interventions

Because we detect evidence of disconfirmation bias and prior attitude effect—essential mechanisms for triggering polarization according to theories of motivated reasoning—our design appears to satisfy most of the necessary conditions that cause participants to "deposit more supportive evidence and affect in memory" (Taber and Lodge 2006, 757). One could contend that the counter-attitudinal arguments provided by GPT-3 are insufficiently strong to trigger the self-defense mechanisms described in previous work. To address this claim, we provide an empirical test of GPT-3's capacity to generate strong arguments in Appendix B.3. We find that while pro arguments generated by GPT-3 are rated as weaker than human-generated arguments, con arguments—the key driver of attitude polarization—generated by GPT-3 are rated to be just as persuasive as human-generated con arguments. Still, even if we grant that GPT-3 cannot generate strong arguments, previous studies have found that *weak* con arguments generate more "refutative thoughts" than strong con arguments (Cacioppo and Petty 1984; Benoit 1987; Benoit and Smythe 2003). Thus, if GPT-3 is underperforming humans in the creation of counter-attitudinal arguments, low-quality con arguments ought to produce *more* attitude polarization than stronger arguments (to the extent that an accumulation of "refutative thoughts" renders it easier to bolster one's attitudes).

That being said, perhaps GPT-3 arguments are not sufficiently confrontational. The deliberative nature of the thought-listing task could also mute reflexive responses to politically incongenial content that might otherwise be detected using more conventional designs. Moreover, there could be heterogeneity in responses to counter-attitudinal information that our design is ill-equipped to detect, given the large number of experimental conditions. Though the average participant does not appear to polarize, there could be politically relevant subgroups who do. We conduct a third experiment that strengthens exposure to counter-attitudinal information by confronting participants with long-form arguments. We also simplify the experimental design by devising a traditional survey experiment where participants are randomly assigned to different pieces of information.

Experiment 3

From December 12, 2022 to December 13, 2022, we recruited 2,000 participants using the online sample provider Lucid.[9] Given concerns about data quality on this platform (Aronow et al. 2020), we created an additional GPT-3 script that assessed the quality of the open-ended responses on the fly and filtered out participants who provided unintelligible responses before treatment assignment.[10] Since our design hinges on legible open-ended responses to produce counter-attitudinal arguments,

[9] The pre-analysis plan for Experiment 3 can be found here: https://aspredicted.org/QRV_BYR

[10] In a pilot study, we compared hand-coded data and the predictions of the data quality script. Appromixately 5% of open-ended responses tagged as "low quality" had legible input, compared to 81% of those tagged "medium quality" and 91% of those tagged "high quality."

these quality checks were necessary. As described in our pre-analysis plan, our analysis omits participants flagged by GPT-3 as "low quality," those who repeated the example issue position in our instructions, and those who did not receive any output from GPT-3. These variables are measured before treatment assignment, and thus, do not bias our estimates. The total API cost was $33.01.

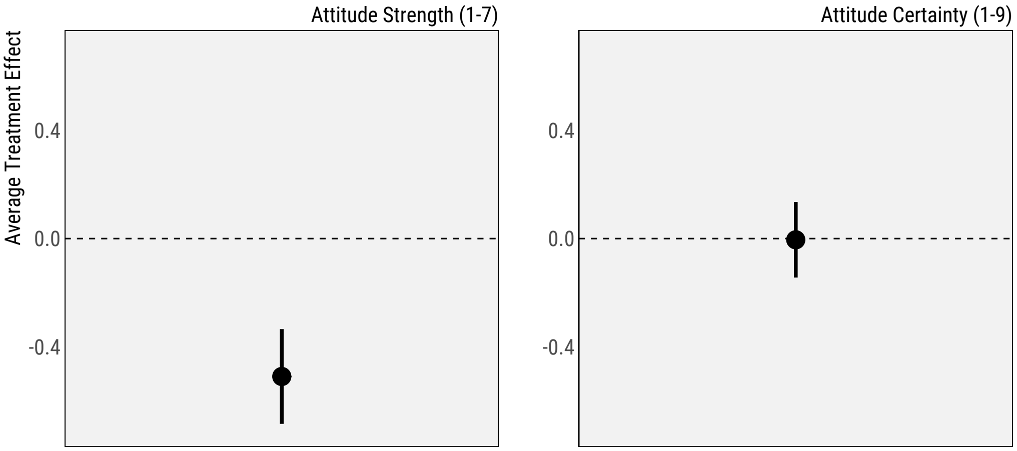
Upon completing the open-ended question, participants responded to a single-question attention check that measured retention of information provided in a news vignette (Kane et al. 2022) and a set of demographic questions. Participants were then randomly shown either (1) one of 4,000 placebo texts drawn from the Porter and Velez (2022) corpus or (2) a tailored counter-argument produced by GPT-3. The placebo texts are news-like vignettes that are approximately a paragraph long, matching the average length of the tailored arguments. Instead of retrieving four pros and cons, we instructed GPT-3 to write a "paragraph-long passionate rebuttal" where the author "strongly disagrees" with the statement provided by the participant (See Appendix C.5 for example arguments).[11]

Following our pre-registration plan, we estimate a covariate-adjusted average treatment effect for attitude strength and certainty, adjusting for age, education, income, partisanship, ideology, and race (white = 1; non-white = 0). We estimate these models using OLS regression with HC2 standard errors. Focusing on the first panel of Figure 5, the covariate-adjusted average treatment effect of counter-attitudinal information on attitude strength is -.510 scale points on a 7-point scale (SE = .09; p < .001).[12] This corresponds to a shift of .26 control-group standard deviation units and is comparable to other effect sizes detected in the persuasion literature (e.g., Broockman and Kalla (2016) detect a shift of .40 scale points in response to their canvassing intervention). An exploratory analysis of treatment effect heterogeneity reveals that effects on attitude strength are consistent across groups varying in open-ended response quality, attentiveness, ideology, and political knowledge. In general, we find that conditional average treatment effects (CATEs) are always negative and statistically significant across the range of moderators. In the case of political knowledge, we find that CATEs are even *more negative* among the most politically sophisticated. The difference between the lower and upper tertile is -.433 scale points (SE = .22; p = .048), suggesting a larger "attitude moderation" effect among those who are more politically knowledgeable. This runs contrary to expectations

[11] The full list of arguments can be found here: bit.ly/3XjLjZC

[12] The ATE (without covariate adjustment) is -.487 scale points (SE = .09; p < .001).

FIGURE 5. Effect of Counter-Attitudinal Information on Attitude Strength and Certainty



Note: This figure presents point estimates and confidence intervals for the covariate-adjusted average treatment effect of counter-attitudinal information on attitude strength and certainty. 95% confidence intervals are presented alongside point estimates.

from the motivated reasoning literature that political sophisticates are especially likely to reject counter-attitudinal information. Turning to certainty, we find that the effect of counter-attitudinal information on certainty is approximately zero ($\hat{ATE}$ = -.005; SE = .07; p = .95).[13] In sum, we find robust evidence of a decrease in attitude strength when participants are exposed to tailored counter-attitudinal information, but are unable to detect evidence of shifts in attitude certainty.

Conclusion

In this paper, we reviewed the literature on attitude polarization and aimed for a critical test of the phenomenon. We exposed participants to tailored counter-attitudinal arguments, measured attitudes using various methods, and primed participants to think in more directional terms. Based on a careful reading of the literature on MR, we sought to create ideal conditions for detecting attitude polarization across two studies. Despite strong theoretical predictions, we did not detect evidence of attitude polarization in either study. In Experiments 1 and 2, we observed small but statistically significant effects in the opposite direction: attitudes became *less* certain after seeing

[13] To minimize multiple comparisons, we pre-registered an analysis of only these two outcomes. However, an exploratory analysis of the single item certainty scale reveals a small negative ATE ($\hat{ATE}$ = -1.708; SE = .87; p = .05).

counter-attitudinal material. The third study simplified the experimental design and exposed participants to longer and more affectively charged arguments. Here, we observed significant decreases in attitude strength, reflecting a moderation process. Despite failing to detect evidence of attitude polarization, we reliably replicated findings such as disconfirmation bias and the prior attitude effect, thus providing these psychological processes with further empirical support.

Our study has important implications for the study of motivated reasoning. If "motivated skepticism" is contingent on strong attitudes, conventional experimental designs might underestimate its prevalence. Attitudes measured using our tailored approach were generally at the upper end of dimensions such as stability, strength, and certainty. Standard measurement approaches involving closed-ended questions may have trouble capturing the nuanced and deeply personal issue positions that define individuals' political orientations. As we discuss in Appendix B.8, while salient issues such as abortion and health care are mentioned frequently by participants, no single issue accounts for more than a quarter of responses. Future research could investigate the extent to which tailored measures of attitudes are better able to capture strong and nuanced attitudes, as well as how they differ from more traditional survey measures.

The use of large-language models (LLMs) and the ability to tailor outcome measures also opens up possibilities for understanding the effectiveness of persuasion when attitudes are deeply entrenched. One could imagine variations of our design that incorporate source cues, test the effectiveness of different argument formats, and assess the durability of persuasive effects. In addition, LLMs can be used to simulate conversations or construct sophisticated chatbots. This could provide opportunities to study the dynamics of persuasion in more social environments. The tailoring of messages to individuals could also yield important insights in domains such as misinformation when targeting rare but socially consequential conspiracy theories and beliefs. Furthermore, the precision and quality of GPT-3 output improves when the base model is fine-tuned using task-specific training data (Brown et al. 2020). We hope to employ the data gathered here to further refine GPT-3 for future studies of political persuasion and information processing.

Motivated reasoning has had a profound impact on political psychology and our larger discipline. In this paper, we consider one of the most troubling implications of motivated reasoning:

attitude polarization. Attitude polarization is troubling because it suggests that deliberation can harden opinions, factual corrections can further reinforce false beliefs, and persuading "true believers" can backfire. Our results suggest that the tendency to increase attitude strength and confidence in the face of counter-attitudinal information may be less common that previously thought, though more work is needed to explore the conditions and moderators of this process. Although we fail to detect evidence of attitude polarization, our tailored approach provides a possible path forward for studying how strong attitudes respond to direct assaults. These studies may disconfirm what we have presented here. We hope that scholars will move in accordance with the evidence.

References

Achen, Christopher H. and Larry M. Bartels (2017). *Democracy for Realists: Why Elections Do Not Produce Responsive Government* (REV - Revised ed.). Princeton University Press.

Aronow, Peter Michael , Joshua Kalla, Lilla Orr, and John Ternovski (2020). Evidence of rising rates of inattentiveness on lucid in 2020.

Bartels, Larry M. (2002, June). Beyond the Running Tally: Partisan Bias in Political Perceptions. *Political Behavior 24*(2), 117–150.

Bayes, Robin , James N. Druckman, Avery Goods, and Daniel C. Molden (2020). When and how different motives can drive motivated political reasoning. *Political Psychology 41*(5), 1031–1052.

Benoit, William L (1987). Argumentation and credibility appeals in persuasion. *Southern Journal of Communication 52*(2), 181–197.

Benoit, William L and Mary Jeanette Smythe (2003). Rhetorical theory as message reception: A cognitive response approach to rhetorical theory and criticism. *Communication Studies 54*(1), 96–114.

Benoît, Jean Pierre and Juan Dubra (2019). Apparent bias: What does attitude polarization show? *International Economic Review 60*(4), 1675–1703. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/iere.12400.

Blumenau, Jack and Benjamin E. Lauderdale (2022, March). The Variable Persuasiveness of Political Rhetoric. *American Journal of Political Science n/a*(n/a). Publisher: John Wiley & Sons, Ltd.

Boninger, David S. , Jon A. Krosnick, and Matthew K. Berent (1995). Origins of Attitude Importance: Self-Interest, Social Identification, and Value Relevance. *Journal of Personality and Social Psychology 68*, 61–80. Place: US Publisher: American Psychological Association.

Broockman, David and Joshua Kalla (2016). Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science 352*(6282), 220–224.

Brown, Tom , Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). Language Models are Few-Shot Learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 1877–1901. Curran Associates, Inc.

Cacioppo, John T and Richard E Petty (1984). The elaboration likelihood model of persuasion. *ACR North American Advances.*

Clifford, Scott , Geoffrey Sheagley, and Spencer Piston (2021). Increasing precision without altering treatment effects: Repeated measures designs in survey experiments. *American Political Science Review 115*(3), 1048–1065.

Clore, Gerald L. and Barbara Baldridge (1968). Interpersonal Attraction: The Role of Agreement and Topic Interest. *Journal of Personality and Social Psychology 9*, 340–346. Place: US Publisher: American Psychological Association.

Coppock, Alexander , Thomas J Leeper, and Kevin J Mullinix (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences 115*(49), 12441–12446.

Ecker, Ullrich KH and Li Chang Ang (2019). Political Attitudes and the Processing of Misinformation Corrections. *Political Psychology 40*(2), 241–260. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pops.12494.

Festinger, Leon (1957). *A Theory of Cognitive Dissonance.* Stanford University Press. Pages: xi, 291.

Gaines, Brian J. , James H. Kuklinski, Paul J. Quirk, Buddy Peyton, and Jay Verkuilen (2007, November). Same Facts, Different Interpretations: Partisan Motivation and Opinion on Iraq. *The Journal of Politics 69*(4), 957–974. Publisher: The University of Chicago Press.

Garrett, R. Kelly , Erik C. Nisbet, and Emily K. Lynch (2013, August). Undermining the Corrective Effects of Media-Based Political Fact Checking? The Role of Contextual Cues and Naïve Theory. *Journal of Communication 63*(4), 617–637.

Gopinath, Mahesh and Prashanth U. Nyer (2009, March). The effect of public commitment on resistance to persuasion: The influence of attitude certainty, issue importance, susceptibility to normative influence, preference for consistency and source proximity. *International Journal of Research in Marketing 26*(1), 60–68.

Guess, Andrew and Alexander Coppock (2020). Does Counter-Attitudinal Information Cause Backlash? Results from Three Large Survey Experiments. *British Journal of Political Science 50*(4), 1497–1515. Edition: 2018/11/05 Publisher: Cambridge University Press.

Haglin, Kathryn (2017, July). The limitations of the backfire effect. *Research & Politics 4*(3), 2053168017716547. Publisher: SAGE Publications Ltd.

Hainmueller, Jens , Jonathan Mummolo, and Yiqing Xu (2019). How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice. *Political Analysis 27*(2), 163–192.

Hart, P. Sol and Erik C. Nisbet (2012, December). Boomerang Effects in Science Communication: How Motivated Reasoning and Identity Cues Amplify Opinion Polarization About Climate Mitigation Policies. *Communication Research 39*(6), 701–723. Publisher: SAGE Publications Inc.

Healy, Andrew J. , Neil Malhotra, and Cecilia Hyunjung Mo (2010, July). Irrelevant events affect voters' evaluations of government performance. *Proceedings of the National Academy of Sciences 107*(29), 12804–12809. Publisher: Proceedings of the National Academy of Sciences.

Houston, David A and Russell H Fazio (1989). Biased Processing as a Function of Attitude Accessibility: Making Objective Judgments Subjectively. *Social cognition 7*(1), 51–66.

Howe, Lauren C and Jon A Krosnick (2017). Attitude Strength. *Annual Review of Psychology 68*(1), 327–351.

Jerit, Jennifer and Jason Barabas (2012, July). Partisan Perceptual Bias and the Information Environment. *The Journal of Politics 74*(3), 672–684. Publisher: The University of Chicago Press.

Kane, John V , Yamil R Velez, and Jason Barabas (2022). Analyze the attentive & bypass bias: Mock vignette checks in survey experiments.

Kraft, Patrick W. , Milton Lodge, and Charles S. Taber (2015, March). Why People "Don't Trust the Evidence": Motivated Reasoning and Scientific Beliefs. *The ANNALS of the American Academy of Political and Social Science 658*(1), 121–133. Publisher: SAGE Publications Inc.

Krosnick, Jon A. (1988, May). Attitude Importance and Attitude Change. *Journal of Experimental Social Psychology 24*(3), 240–255.

Krosnick, Jon A. (1990, March). Government policy and citizen passion: A study of issue publics in contemporary America. *Political Behavior 12*(1), 59–92.

Krosnick, Jon A. and Richard E. Petty (1995). Attitude Strength: An Overview. In *Attitude strength: Antecedents and Consequences*, Ohio State University series on attitudes and persuasion, Vol. 4., pp. 1–24. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Kuhn, Deanna and Joseph Lao (1996, March). Effects of Evidence on Attitudes: Is Polarization the Norm? *Psychological Science 7*(2), 115–120. Publisher: SAGE Publications Inc.

Kuklinski, James H. , Paul J. Quirk, Jennifer Jerit, David Schwieder, and Robert F. Rich (2000, August). Misinformation and the Currency of Democratic Citizenship. *The Journal of Politics 62*(3), 790–816. Publisher: The University of Chicago Press.

Kunda, Ziva (1987). Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories. *Journal of Personality and Social Psychology 53*, 636–647. Place: US Publisher: American Psychological Association.

Kunda, Ziva (1990). The Case for Motivated Reasoning. *Psychological bulletin 108*(3), 480. Publisher: American Psychological Association.

Lavine, Howard , Eugene Borgida, and John L. Sullivan (2000, March). On the Relationship Between Attitude Involvement and Attitude Accessibility: Toward a Cognitive-Motivational Model of Political Information Processing. *Political Psychology 21*(1), 81–106. Publisher: John Wiley & Sons, Ltd.

Little, Andrew T (2021). Detecting Motivated Reasoning. *URL: osf. io/b8tvk*.

Lodge, Milton and Charles S. Taber (2005, June). The Automaticity of Affect for Political Leaders, Groups, and Issues: An Experimental Test of the Hot Cognition Hypothesis. *Political Psychology 26*(3), 455–482. Publisher: John Wiley & Sons, Ltd.

Lord, Charles G , Lee Ross, and Mark R Lepper (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology 37*(11), 2098. Publisher: American Psychological Association.

McHoskey, John W. (1995, October). Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization. *Basic and Applied Social Psychology 17*(3), 395–409. Publisher: Routledge.

Miller, Arthur G. , John W. McHoskey, Cynthia M. Bane, and Timothy G. Dowd (1993). The Attitude Polarization Phenomenon: Role of Response Measure, Attitude Extremity, and Behavioral Consequences of Reported Attitude Change. *Journal of Personality and Social Psychology 64*, 561–574. Place: US Publisher: American Psychological Association.

Mullinix, Kevin J , Thomas J Leeper, James N Druckman, and Jeremy Freese (2015). The generalizability of survey experiments. *Journal of Experimental Political Science 2*(2), 109–138.

Munro, Geoffrey D. and Peter H. Ditto (1997, June). Biased Assimilation, Attitude Polarization, and Affect in Reactions to Stereotype-Relevant Scientific Information. *Personality and Social Psychology Bulletin 23*(6), 636–653. Publisher: SAGE Publications Inc.

Nisbett, Richard E. and Lee Ross (1980). *Human Inference: Strategies and Shortcomings of Social Judgment.* Englewood Cliffs, NJ, USA: Prentice-Hall.

Nyhan, Brendan , Ethan Porter, Jason Reifler, and Thomas J. Wood (2020, September). Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability. *Political Behavior 42*(3), 939–960.

Nyhan, Brendan and Jason Reifler (2010, June). When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior 32*(2), 303–330.

Nyhan, Brendan and Jason Reifler (2015, January). Does correcting myths about the flu vaccine work? An experimental evaluation of the effects of corrective information. *Vaccine 33*(3), 459–464.

Petrocelli, John V , Zakary L Tormala, and Derek D Rucker (2007). Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of personality and social psychology 92*(1), 30.

Pomerantz, Eva M. , Shelly Chaiken, and Rosalind S. Tordesillas (1995). Attitude Strength and Resistance Processes. *Journal of Personality and Social Psychology 69*, 408–419. Place: US Publisher: American Psychological Association.

Porter, Ethan and Yamil R. Velez (2022). Placebo Selection in Survey Experiments: An Agnostic Approach. *Political Analysis 30*(4), 481–494. Edition: 2021/06/14 Publisher: Cambridge University Press.

Redlawsk, David P. (2002, November). Hot Cognition or Cool Consideration? Testing the Effects of Motivated Reasoning on Political Decision Making. *Journal of Politics 64*(4), 1021–1044. Publisher: John Wiley & Sons, Ltd.

Ryan, Timothy J and J Andrew Ehlinger (2019). Issue Publics: Fresh Relevance for an Old Concept. Working paper presented at the Annual Meeting of the American Political ....

Schaffner, Brian F and Cameron Roche (2017). Misinformation and Motivated Reasoning: Responses to Economic News in a Politicized Environment. *The Public Opinion Quarterly 81*(1), 86–110. Publisher: [Oxford University Press, American Association for Public Opinion Research].

Shapiro, Robert Y. and Yaeli Bloch-Elkon (2008, January). Do the Facts Speak for Themselves? Partisan Disagreement as a Challenge to Democratic Competence. *Critical Review: A Journal of Politics and Society 20*(1-2), 115–139. Publisher: Routledge.

Swire-Thompson, Briony , Joseph DeGutis, and David Lazer (2020, September). Searching for the Backfire Effect: Measurement and Design Considerations. *Journal of Applied Research in Memory and Cognition 9*(3), 286–299.

Taber, Charles S and Milton Lodge (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science 50*(3), 755–769.

Tesser, Abraham and Christopher Leone (1977, July). Cognitive Schemas and Thought as Determinants of Attitude Change. *Journal of Experimental Social Psychology 13*(4), 340–356.

U.S. Policy Agendas Project (2019). Topics Codebook.

Vidigal, Robert and Jennifer Jerit (2022, November). Issue Importance and the Correction of Misinformation. *Political Communication 39*(6), 715–736. Publisher: Routledge.

Walter, Nathan , Jonathan Cohen, R Lance Holbert, and Yasmin Morag (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication 37*(3), 350–375.

Weeks, Brian E. (2015, August). Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation. *Journal of Communication 65*(4), 699–719.

Wood, Thomas and Ethan Porter (2019). The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior 41*(1), 135–163. Publisher: Springer.

Zhou, Jack (2016, September). Boomerangs versus Javelins: How Polarization Constrains Communication on Climate Change. *Environmental Politics 25*(5), 788–811.

Zuwerink Jacks, Julia R. and Patricia G. Devine (1996). Attitude Importance and Resistance to Persuasion: It's Not Just the Thought That Counts. *Journal of Personality and Social Psychology 70*, 931–944. Place: US Publisher: American Psychological Association.

Appendix

A Pre-Registered Analyses

A.1 Full Model Results (Experiment 1)

TABLE A1 . Effects of Information and Motivation Primes on Attitudes (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.845***	97.117***
	(0.024)	(0.164)
Pre-Treatment Strength	0.272***	0.043
	(0.064)	(0.189)
Pre-Treatment Certainty	0.182***	6.133***
	(0.033)	(0.466)
Con	-0.015	-0.025
	(0.038)	(0.359)
Mixed	-0.022	-0.004
	(0.035)	(0.230)
Directional Prime	0.012	0.053
	(0.033)	(0.213)
Con × Directional Prime	-0.007	-0.123
	(0.051)	(0.469)
Mixed × Directional Prime	-0.004	-0.900*
	(0.049)	(0.429)
N	1768	1772
R^2	0.438	0.718

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: OLS regression with HC2 robust standard errors.

A.2 Full Model Results (Experiment 2)

TABLE A2 . Effects of Information and Motivation Primes on Attitudes (Experiment 2)

	Attitude Strength (1-7)	Certainty (0-100)	Certainty Score (1-9)
(Intercept)	5.951***	86.923***	7.703***
	(0.077)	(0.867)	(0.056)
Pre-Treatment Strength	0.804***	1.511**	0.132**
	(0.057)	(0.548)	(0.042)
Pre-Treatment Certainty	0.140**	9.973***	
	(0.046)	(0.651)	
Con	0.104	−0.396	−0.127+
	(0.103)	(0.990)	(0.067)
Mixed	0.077	−1.152	−0.184**
	(0.087)	(1.047)	(0.069)
Directional Prime	0.130	1.076	0.012
	(0.086)	(0.871)	(0.064)
Weak Attitude	−0.123	−1.613	0.041
	(0.133)	(1.709)	(0.117)
Con × Directional Prime	−0.190	−0.334	−0.023
	(0.128)	(1.236)	(0.098)
Mixed × Directional Prime	−0.113	−0.181	0.112
	(0.113)	(1.294)	(0.096)
Con × Weak Attitude	−0.321+	−3.357	−0.391*
	(0.165)	(2.193)	(0.165)
Mixed × Weak Attitude	−0.456**	−2.777	−0.199
	(0.155)	(2.162)	(0.160)
Directional Prime × Weak Attitude	−0.384*	−0.798	−0.253
	(0.161)	(2.169)	(0.159)
Con × Directional Prime × Weak Attitude	0.378+	−1.356	0.511*
	(0.223)	(3.071)	(0.230)
Mixed × Directional Prime × Weak Attitude	0.541*	1.721	0.155
	(0.216)	(2.947)	(0.230)
Pre-Treatment Certainty Scores			0.925***
			(0.050)
N	2338	2331	2357
R^2	0.485	0.413	0.456
Clustered Standard Errors	by: id	by: id	by: id

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: OLS regression with CR2 robust standard errors.

A.3 Full Model Results (Experiment 3)

TABLE A3 . Effects of Information on Attitudes (Experiment 3)

[a]

	Attitude Strength (1-7)	Attitude Strength (1-7)	Attitude Certainty (1-9)	Attitude Certainty (1-9)
(Intercept)	6.043***	6.038***	7.630***	7.630***
	(0.056)	(0.056)	(0.053)	(0.052)
Treatment	−0.487***	−0.510***	0.015	−0.005
	(0.089)	(0.089)	(0.073)	(0.071)
Age		0.152**		0.333***
		(0.049)		(0.037)
Education		0.056		0.033
		(0.050)		(0.040)
Income		0.054		0.122***
		(0.047)		(0.037)
Ideology		−0.072		−0.077
		(0.059)		(0.050)
White		0.104*		0.081*
		(0.048)		(0.040)
Partisanship		−0.074		0.005
		(0.057)		(0.048)
N	1891	1848	1895	1852
R^2	0.016	0.034	0.000 02	0.065

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

[a] OLS regression with HC2 robust standard errors.

A.4 Validation Tests (Experiment 1)

We assess whether we are able to detect disconfirmation bias by regressing the share of denigrating responses to the GPT-3-produced arguments and total response time for the thought-listing task on information condition indicators, priming condition indicators, and their interaction. We measure the share of denigrating responses by asking crowdsourcing workers on Amazon Mechanical Turk to count the number of denigrating responses for four sets of argument-response pairs. For each respondent, we divide the number of denigrating responses by the total number of responses provided ($\bar{x} = .17$). We detect clear evidence of disconfirmation bias. Roughly 9% of thought-listing responses in the "Pro" condition rejected or dismissed the arguments that were presented, whereas the share of denigrating comments was 8pp higher (SE = 1pp) in the "Mixed" condition and 16pp higher (SE = 2pp) in the "Con" condition. A larger (smaller) share of denigrating comments were observed when "Con" ("Mixed") information was paired with a directional prime. However, the difference-in-difference estimates for both are not statistically significant. Moving on to timer data, we find evidence of slower response times for those in "Mixed" and "Con" conditions. These correspond to approximately 5% (SE = .008; $p < .001$) and 6% decreases (SE = .008; $p < .001$) in response times, respectively. Turning to the interaction model, those in the "Pro" condition who receive a directional prime spend approximately 3% less time on the thought-listing task than those who receive an accuracy prime (SE = .012; $p < .001$). However, we fail to find evidence that the directional prime conditions the effects of the "Con" and "Mixed" conditions on response time. Taken together, the evidence is broadly consistent with research on disconfirmation bias finding that people are more likely to denigrate and expend cognitive resources on counter-attitudinal versus pro-attitudinal information.

TABLE A4 . The Effect of Information Conditions on Share of Denigrating Comments and Response Time Data

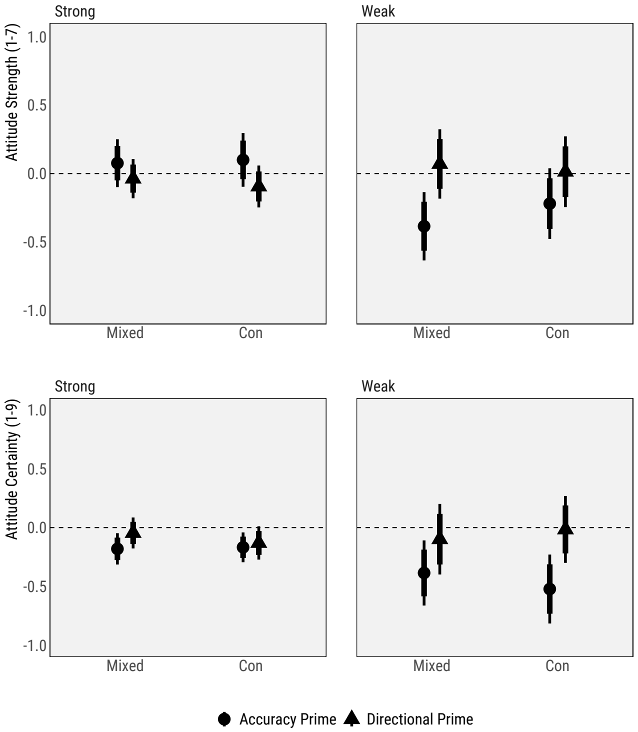
	Denigrating Share	Denigrating Share	Log Timer	Log Timer
(Intercept)	0.092***	0.086***	1.524***	1.542***
	(0.008)	(0.012)	(0.006)	(0.009)
Information (Mixed)	0.083***	0.095***	0.048***	0.040**
	(0.013)	(0.019)	(0.008)	(0.012)
Information (Con)	0.158***	0.145***	0.059***	0.050***
	(0.019)	(0.025)	(0.008)	(0.012)
Directional Prime		0.012		−0.031**
		(0.017)		(0.012)
Information (Con) × Directional Prime		0.030		0.013
		(0.037)		(0.016)
Information (Mixed) × Directional Prime		−0.025		0.010
		(0.026)		(0.017)
N	1729	1729	1782	1782

Note: Statistical significance levels: + p<.10; * p< .05; ** p<.01; *** p<.001

A.5 Differences Between Strong and Weak Attitudes (Experiment 2)

As described in the manuscript, we do not detect significant shifts in attitudes across information conditions for the strong attitudes. However, when we focus on attitudes toward more peripheral issues, we observe some mixed evidence of persuasion. Those in the "Con" condition who were assigned to an accuracy prime score .22 scale points lower (SE = .13; p = .10) on attitude strength than those in the "Pro" condition. Examining those in the "Mixed" condition, participants score about .38 scale points lower on attitude strength than those in the "Pro" condition (SE = .13; p = .003) when they receive an accuracy prime. The two estimates correspond to .17 and .3 standard deviation unit shifts on the outcome variable. Though we detect shifts in certainty for the strong attitudes, effects on weak attitudes are generally larger. Certainty scores are .02 (SE = .15) and .10 (SE = .15) scale points lower for those in the "Con" and "Mixed" conditions relative to the "Pro" condition when respondents receive a directional prime. When they are primed to be accurate, difference-in-means estimates for the "Con" and "Mixed" condition vis-a-vis the "Pro" condition drop to -.52 (SE = .15) and -.39 (SE = .14) scale points, respectively. These two estimates are equivalent to .26 and .2 standard deviation unit shifts in the outcome. Overall, GPT-3 is capable of producing persuasive effects, especially when focusing on less crystallized attitudes.

FIGURE A1 . Effects on Attitude Strength and Certainty Across Information and Issue Strength Conditions

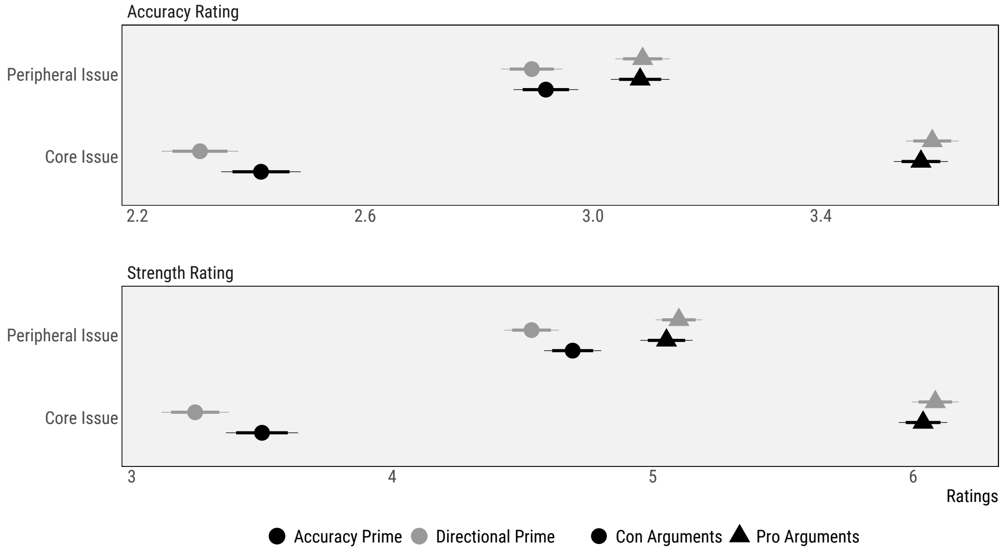


Note: This figure presents point estimates and confidence intervals for attitude strength and certainty across information and motivation conditions with facets defined by issue attitude strength. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands.

A.6 Prior Attitude Effect (Experiment 2)

Following the attitude strength and certainty measures within each trial, participants were asked to rate the full set of eight arguments provided by GPT-3 on factual accuracy and argument strength. Factual accuracy was measured using a traditional four-point ordinal scale ranging from "not at all accurate" to "very accurate." Argument strength was measured using a seven-point item ranging from "very weak" to "very strong." Comparing the differences in mean ratings between pro arguments for the strong versus peripheral issue, participants generally rate pro arguments .98 (SE = .06; p < .001) scale points lower on argument strength and .50 (SE = .03; p < .001) scale points lower on factual accuracy when responding to arguments concerning peripheral versus core issues. Focusing on con arguments, ratings for argument strength and factual accuracy are 1.30 (SE = .07; p < .001) and .56 (SE = .04; p < .001) scale points higher, respectively, when peripheral versus core issues are considered. For the core issue, the gap between pro and con arguments in accuracy and argument strength is 1.22 scale points (SE = .03; p < .001) and 2.69 scale points (SE = .06; p < .001). This gap shrinks by 1.04 scale points (SE = .04; p < .001) and 2.23 scale points (SE = .08; p < .001) when individuals are considering arguments related to the peripheral issue. As a validation of our motivational primes, we also find evidence that accuracy primes generally shift accuracy and strength ratings upward by .12 (SE = .057; p = .035) and .27 (SE = .11; p = .016) scale points, respectively. In sum, we find evidence that argument strength and factual accuracy are conditional on whether arguments relate to core or peripheral issues.

FIGURE A2 . Argument Strength and Accuracy Ratings Across Core and Peripheral Issues

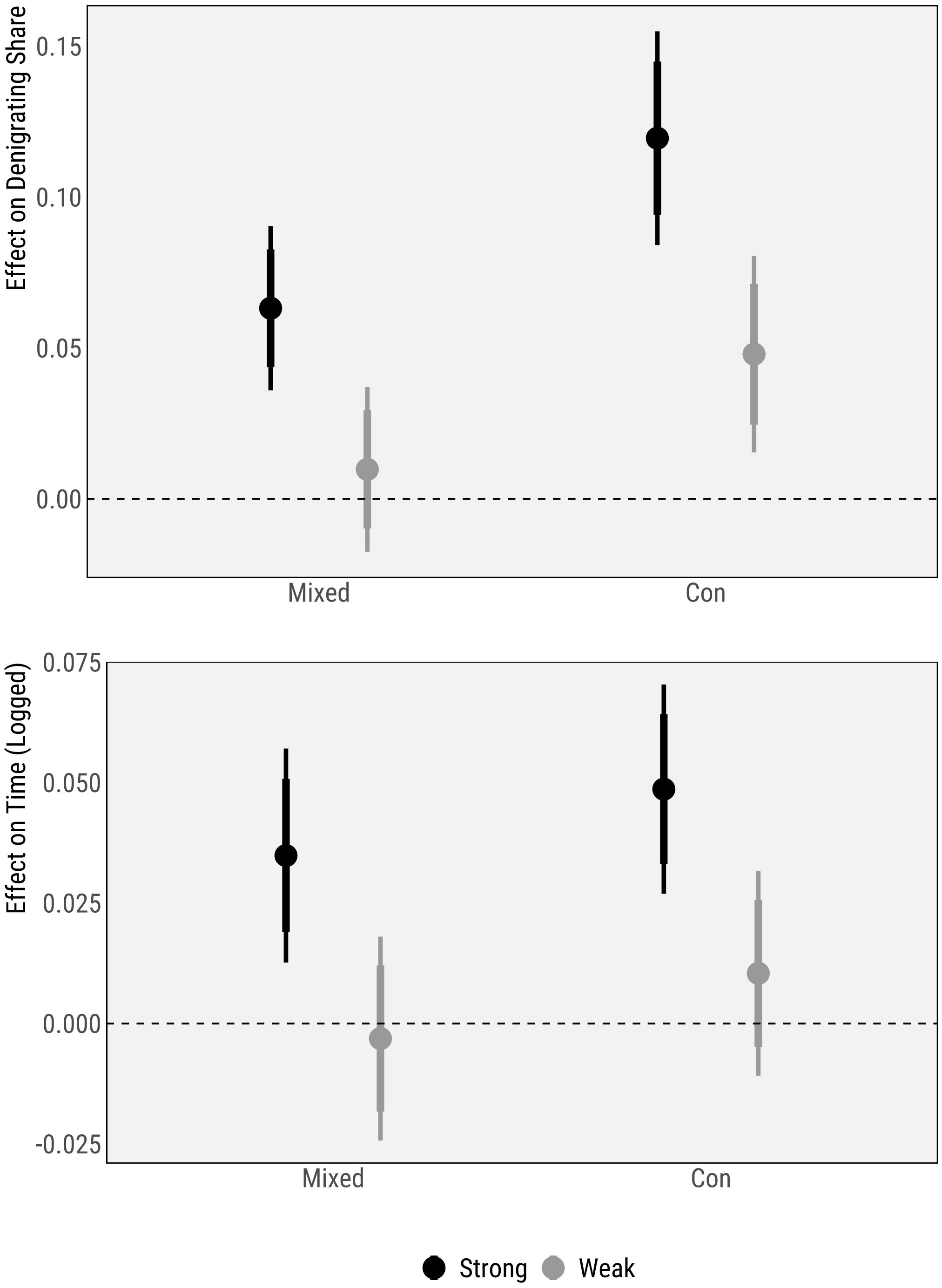


Note: This figure presents point estimates and confidence intervals for accuracy and strength ratings across information, motivation, and issue strength conditions. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands.

A.7 Disconfirmation Bias (Experiment 2)

As shown in Figure A3 , counter-attitudinal information generally performs differently across “weak” and “strong” attitudes, such that participants are more likely to reject or disagree with “Mixed” and “Con” arguments when they respond to a core issue versus a more peripheral issue. Focusing on the top panel, those in the “Mixed” condition produce a 6pp greater share of denigrating comments (SE = 1pp; p < .001) relative to those in the “Pro” condition when the arguments concerns a strong attitude. This number drops to 1pp (SE = 1pp; n.s.) when the arguments involve a weak attitude. The difference between these two estimates is significant (*d* = 5pp; SE = 2pp; p = .007). Turning to the “Con” condition, the share of denigrating comments increases by 12pp (SE = 2pp; p < .001) when the arguments target a strong attitude, relative to 5pp (SE = 2pp; p = .003) for the weak attitude. This difference is statistically significant (*d* = 7pp; SE = 2pp; p < .001). Data on response times evince a similar pattern. Since response times are logged, estimates are interpreted on a percentage (not percentage point) basis. Those responding to “Mixed” information spend approximately 3.4% longer on arguments (SE = 1pp; p <.001) relative to the “Pro” condition when these arguments describe a strong attitude. This number drops to -.03% (SE = 1pp; n.s.) for the weak attitude. This difference is statistically significant (*d* = .038; SE = .016; p = .01). Those responding to four cons spend 4.8% longer on the thought-listing task (SE = 1pp; p < .001) when the arguments are attitudinally inconsistent, whereas this number is 1% (SE = .01; n.s.) when weak attitudes are challenged. This difference is also statistically significant (*d* = .038; SE = .016; p = .01). In sum, we find evidence that we are activating the mechanisms identified by previous research. However, even in doing so, we fail to detect evidence of attitude polarization.

FIGURE A3 . Disconfirmation Bias: Denigrating Comments and Response Times



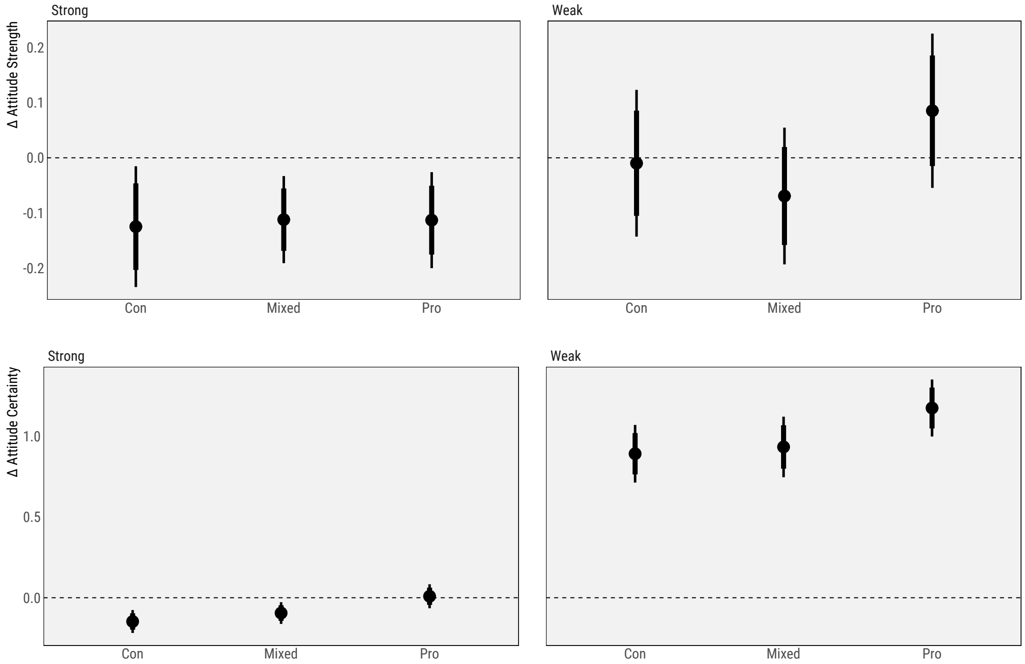
Note: This figure presents point estimates and confidence intervals for models examining the share of denigrating comments and response times across information, motivation, and issue strength conditions. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands.

B Exploratory Analyses

B.1 Change Score Analysis (Experiment 2)

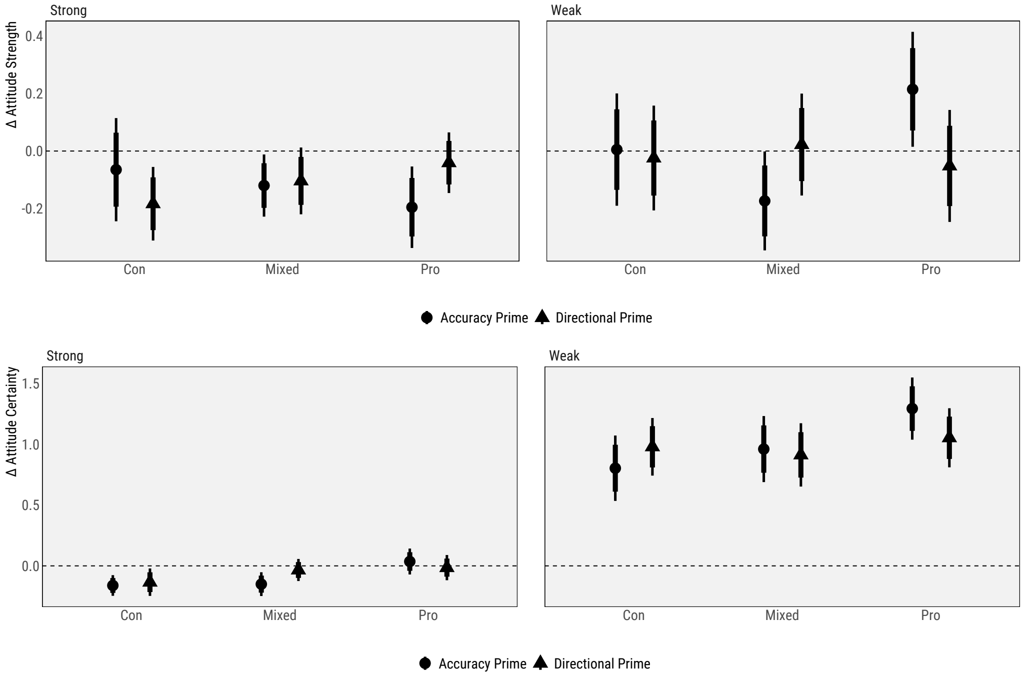
As shown in Figure B1 , we fail to detect relative increases in attitude strength or certainty for those in the counter-attitudinal conditions. Focusing on attitude strength, we generally detect moderation across the board for the "strong attitude," with those exposed to "Con," "Mixed," and "Pro" information reporting weaker attitudes after the informational and motivational interventions. Estimates for the "Con," "Mixed," and "Pro" condition are -.089 (SE = .049), -.118 (SE = .0427), and -.11 (SE = .056) scale points, respectively. Changes in attitude strength for the weak condition are muted and not statistically discernible from zero, ranging from -.006 to .052 scale points. Turning now to attitude certainty, certainty shifts by -.146 in the "Con" condition, -.08 in the "Mixed" condition, and .023 in the "Pro" condition when assessing the "strong attitude." The difference in Δ attitude certainty between the "Pro" and "Con" condition (*d* = .16; SE = .06; p = .003) is significant. Finally, turning to the "weak attitude," we observe significant increases in certainty for all informational conditions. Certainty scores increase by .94, .91, and 1.17 scale points in the "Con," "Mixed," and "Pro" conditions, respectively. The change in the "Pro" condition is larger than the counterattitudinal conditions. However, these differences are not statistically significant. Taken together, we detect evidence of moderation with respect to strong attitudes and increases in certainty for weak attitudes. For the latter, increases in certainty occur across the board, which is consistent with a possible learning process whereby exposure to any new information causes an increase in certainty. An additional analysis that examines conditional means across the motivational conditions fails to recover evidence of attitude polarization.

FIGURE B1 . Change Score Analysis of Attitude Strength and Certainty Across Information and Issue Strength Conditions



Note: This figure presents point estimates and confidence intervals for changes in attitude strength and certainty across information conditions with facets defined by issue attitude strength. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands.

FIGURE B2 . Change Score Analysis of Attitude Strength and Certainty Across Information, Motivation, and Issue Strength Conditions

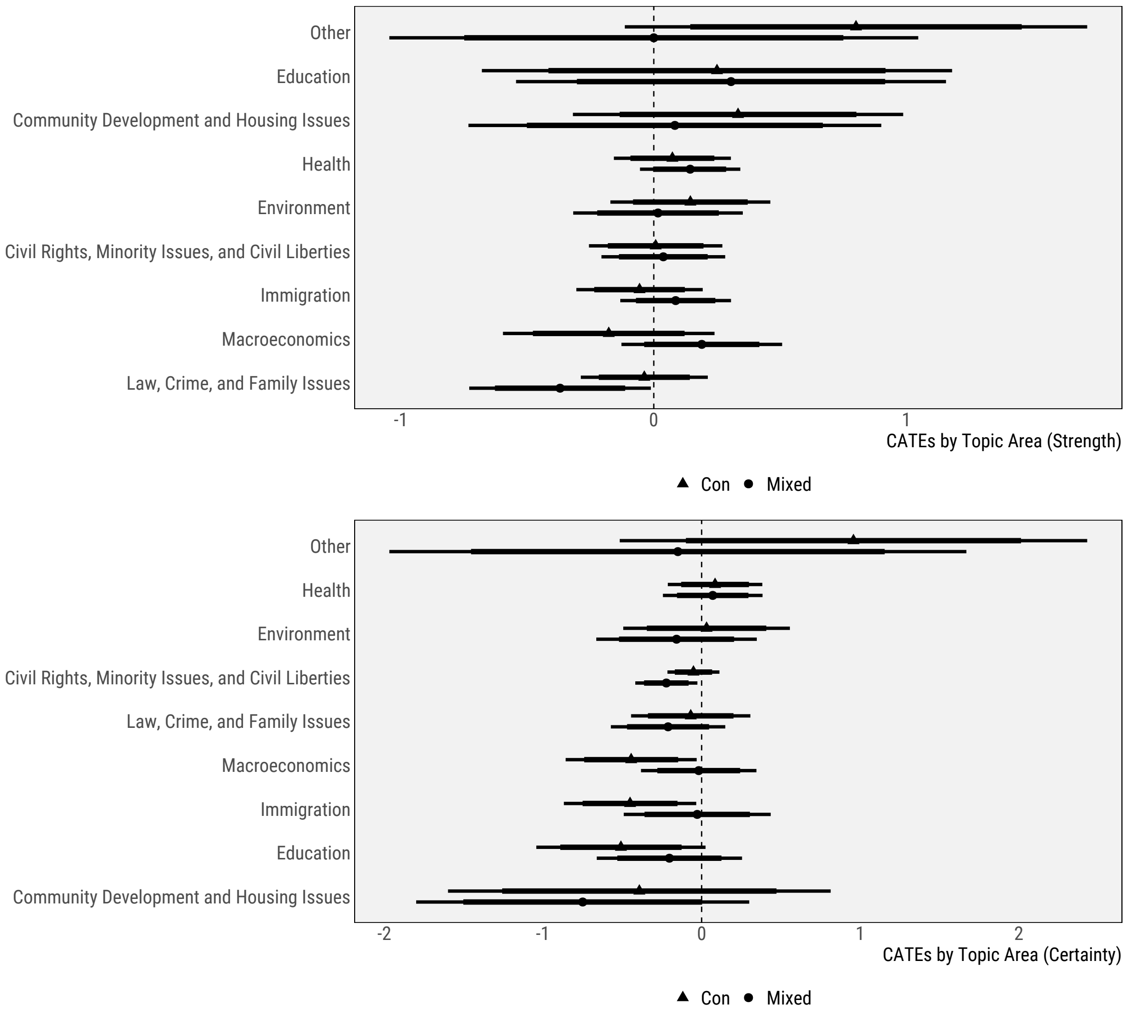


Note: This figure presents point estimates and confidence intervals for changes in attitude strength and certainty across information conditions with facets defined by issue attitude strength and motivation conditions. 84% confidence intervals are used to facilitate visual detection of significant group differences (thick bands). 95% confidence intervals are presented using thin bands.

B.2 Heterogeneous Treatment Effects (Experiment 2)

We assess treatment effect heterogeneity across different issue topics. To define issue topics, we rely on the Comparative Agendas Project coding scheme, which is comprised of 21 major political topics and 220 subtopics. When calculating CATEs and making inferences about issue areas, we rely on major topics, given the small number of participants who write about the same subtopic. Even then, the median number of responses per major issue topic is 23, leaving us with imprecise estimates. Consistent with the ATE estimate, CATEs on the attitude strength measure are small across issue topics, especially when focusing on issue topics with larger subsamples. Due to the lack of precision, the vast majority of effects are not statistically discernible from zero. The same is true for certainty. Still, we fail to detect evidence of attitude polarization when examining relevant subgroups.

FIGURE B3 . CATEs on Attitude Strength by Issue Topic



Note: This figure presents point estimates for conditional average treatment effects by issue topic. CATEs are estimated using a Bayesian random-effects model with varying intercepts and slopes. 95% CIs

B.3 Are GPT-3 arguments persuasive? (Experiment 2)

One concern with GPT-3 is that it is incapable of matching the persuasive strength of arguments created by humans, which explains why participants did not respond strongly to the information conditions. This potential issue does not explain our pattern of findings across the three studies. First, we address this concern directly by conducting a third experiment that exposes participants to longer and more affectively charged arguments. In this experiment, we observe substantively large shifts in attitude strength (approximately .50 of a scale point on a 7-point scale). These effect sizes are comparable to effect sizes detected in other studies of persuasion in political science (.40 of a scale point; Broockman and Kalla (2016)). Second, previous scholars have noted that attitude polarization can arise when counter-attitudinal arguments deposit more "refutative thoughts" in memory. Assuming this mechanism is valid, argument strength may have an inverse relationship with attitude polarization, given that people can more easily refute weaker arguments. If anything, strong arguments can render it more difficult to detect attitude polarization because these arguments are more challenging to dispute and counter-argue.

However, given that other scholars may use these tools for purposes other than detecting attitude polarization, we carried out a descriptive study of argument strength by recruiting 200 human raters on CloudResearch Connect and asking them to rate ten pairs of arguments produced by GPT-3 and humans. Blumenau and Lauderdale (2022, 11-13) show in their validation experiment that asking respondents to rate the persuasiveness of arguments, while different from measuring the degree to which those arguments actually move respondents' beliefs, is nonetheless highly informative about the relative persuasiveness of arguments, as the outcomes of these two methods often closely correlate. For the corpus of human-produced arguments, we relied on Kialo, a collaborative website that allows users to map out the structure of political arguments. On Kialo, a claim is posted (e.g., "the death penalty should be abolished") and users upload supporting and opposing arguments. Users can rate arguments based on strength, leave comments, and provide additional supporting evidence for those main arguments.

Our descriptive study of argument strength utilized a set of 21 arguments about political and policy topics. We adopted the following procedure to generate this set: First, using Experiment 2 data for which we previously coded CAP subtopics, we drew one random observation from each of the 66 unique subtopics. Each observation included one respondent's self-report of a deeply held issue position. For each position, we scoured Kialo's forums for a corresponding discussion prompt. We ended up with 21 such claims. If the pro/con position of the Kialo prompt was opposite that of our survey respondent, we reversed the Kialo position to match the respondent's, and correspondingly flipped the labels on Kialo's pro and con arguments (to con and pro, respectively). Since our data includes four pro and four con arguments generated by GPT-3 for each open-ended response, we similarly selected four pro and four con arguments from each Kialo discussion thread. To select Kialo arguments, we first scraped the four strongest-rated arguments on the main branch of each discussion thread. If the main arguments numbered fewer than four, we selected the strongest-rated supporting argument under each main argument until we reached four. We removed citations and hyperlinks from Kialo arguments to mitigate source cue effects but kept appeals to descriptive statistics and maintained the original length of the arguments, even though these may render such arguments more persuasive than the single-sentence arguments generated by GPT-3.

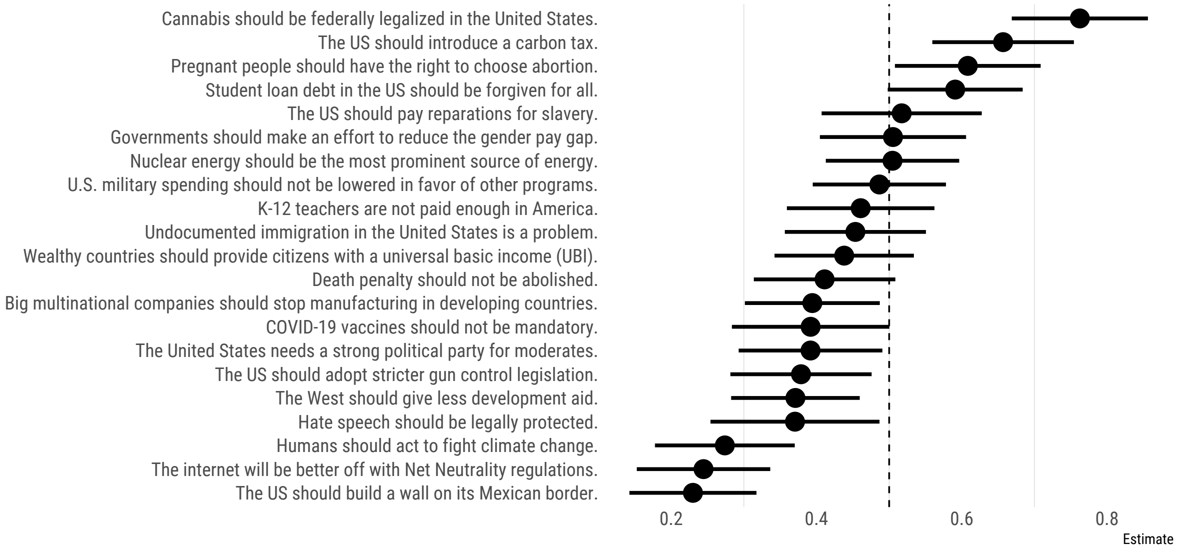
Averaging over the 21 claims, GPT-3 arguments were rated as stronger than Kialo arguments 45% of the time (SE = 1%). This is impressive, given that these arguments were generated on the fly, whereas Kialo arguments are refined and edited by human volunteers. We now examine differences between GPT-3 and human arguments by issue position (e.g., pro and con). 50% of con arguments generated by GPT-3 were rated as superior to Kialo arguments, whereas 40% of pro arguments generated by GPT-3 ”beat out” Kialo arguments. The difference between pro and con arguments is negative and significant (*d* = -9%; SE = 2.3%; p < .001), suggesting that GPT-3 arguments slightly underperformed human arguments when pro arguments were generated. Still, given the importance of con arguments within the context of motivated reasoning, these findings are reassuring; GPT-3 generates con arguments that are rated to be just as persuasive as human arguments. Turning now to variation across issue areas, GPT-3 arguments were rated as more persuasive than human arguments in the domains of cannabis legalization, abortion rights, and carbon taxes, while being rated as less persuasive than human arguments in areas such as climate change, Net Neutrality, and the border wall. For most issues, GPT-3 arguments were rated slightly lower than human arguments, with estimates hovering between 40 and 50%. Overall, we provide evidence that GPT-3 arguments are at least comparable in strength to human arguments, with GPT-3 arguments outperforming human arguments in certain contexts.

TABLE B1 . Comparing GPT-3 and Kialo Arguments

	Model 1	Model 2
(Intercept)	0.450	0.496
	(0.012)	(0.016)
Pro Arguments		−0.093
		(0.024)
N.	1996	1996
R2	2×10^{-14}	0.009
Std.Errors	by: participant	by: participant

Note: Standard errors are clustered by participant.

FIGURE B4 . Comparison between GPT-3 and Human Arguments by Issue

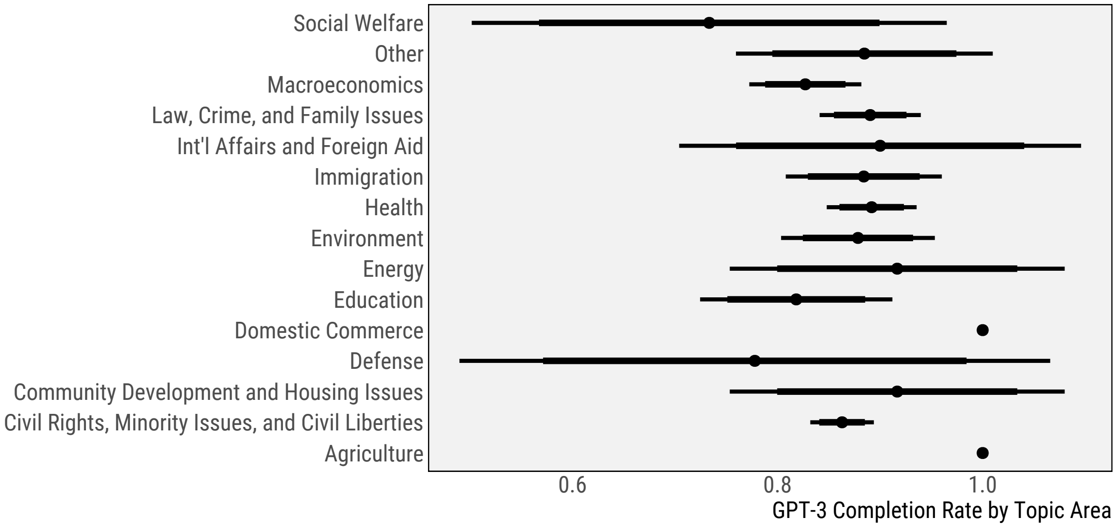


Note: This figure presents point estimates for the proportion of GPT-3 arguments that are rated as stronger than Kialo arguments by issue. 95% (thin bar) CIs are presented alongside point estimates.

B.4 GPT-3 Completion by Issue Topic (Experiment 2)

Although GPT-3 failures occurred a small percentage of the time, these errors occurred before the treatment was administered. As discussed in the paper, participants could be assigned to either (1) four pros, (2) four cons, or (3) two pros and two cons from a balanced list of eight GPT-3 generated pros and cons. This list was produced before treatment assignment, and in cases where GPT-3 could not produce a list of pros and cons, these respondents were flagged in the survey software (i.e., generic_flag = 1) and provided with a generic set of pros and cons discussing government intervention. As we note in our pre-registration plans, we exclude these participants because they did not receive a tailored response. Since all of this occurs before treatment assignment, there is no risk of confounding. However, these errors could affect the representativeness of our estimated treatment effects if certain topics are less likely to receive a tailored response. Though there is variation across issue topics with respect to successful GPT-3 output, differences are small in magnitude, and we fail to reject the null hypothesis in an F-test assessing equality of issue topics with respect to GPT-3 completions ($F(14, 1310) = .79$, p = .68).

FIGURE B5 . Proportion of Valid GPT Completions by Issue Topic



Note: This figure presents point estimates for the proportion of valid GPT-3 responses by issue topic. 84% (thick bar) and 95% (thin bar) CIs are presented alongside point estimates.

TABLE B2 . F-test assessing equality of issue topics

	df	Sum Sq.	Mean Sq.	F	p
Issue Topic	14	1.29	0.09	0.79	0.6830
Residuals	1296	151.09	0.12		

B.5 Certainty Analysis (Experiment 3)

To minimize the possibility of multiple testing, we pre-registered an analysis of only two outcomes (strength and multi-item certainty). However, we also included a single-item measure of certainty in our survey. As an exploratory analysis, we assess if this item was also moved by the intervention. We find evidence that exposure to counter-attitudinal information reduced certainty on this scale by about 1.7 scale points on a 0-100 scale (p = .05, adjusting for the pre-registered set of covariates used in the main models). This corresponds to a shift of approximately .09 control-group standard deviation units.

	Model 1	Model 2
(Intercept)	88.012***	87.945***
	(0.613)	(0.605)
Treatment	-1.527+	-1.708*
	(0.884)	(0.870)
Age		3.195***
		(0.465)
Education		0.276
		(0.487)
Income		0.951*
		(0.465)
Ideology		-0.872
		(0.589)
White		0.281
		(0.476)
Partisanship		0.277
		(0.604)
N	1815	1773
R^2	0.002	0.038

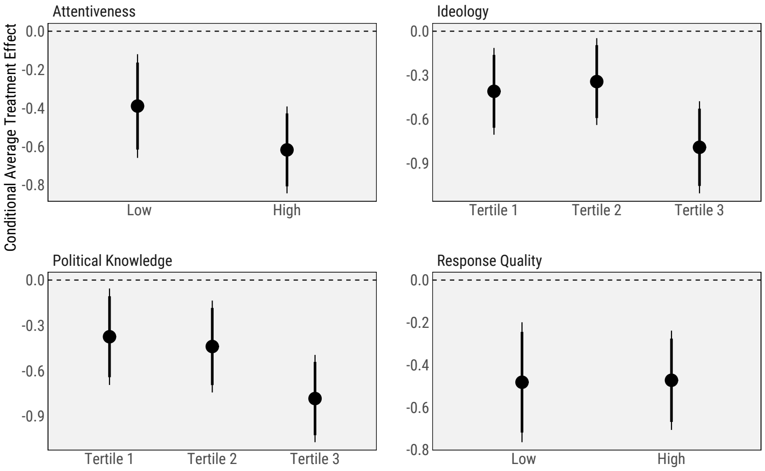
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Covariates are mean centered.

B.6 Robustness of Effects (Experiment 3)

We assess conditional effects across levels of attentiveness, ideology, political knowledge, and response quality. Attentiveness is measured using a question that assesses the retention of information provided in a paragraph-long news vignette. As Kane et al. (2022) show, this method tends to predict longer survey duration, higher factual manipulation check passage, and larger treatment effects. Ideology is measured using a five-point scale. Political knowledge is measured using four questions about American politics (i.e., majority party in the House, most conservative party, size of veto-proof majority, knowledge of Supreme Court's function, former Vice President identification). Finally, response quality is measured using a GPT-3 script ("low" represents the medium quality response, given that we filtered out participants who provided low quality responses). We break the continuous and ordinal variables into tertiles to avoid model dependence issues with linear multiplicative interaction models (Hainmueller et al. 2019) and estimate CATEs using the pre-registered models presented in the manuscript (i.e., covariate-adjusted average treatment effects with HC2 robust standard errors). We detect limited treatment effect heterogeneity across subgroups, but conditional effects generally increase with attentiveness, political knowledge, and conservative self-identification. In all cases, we detect negative point estimates, such that individuals moderate when exposed to counter-attitudinal information.

FIGURE B6 . Individual-level CATEs



Note: This figure presents point estimates for conditional average treatment effects by attentiveness, response quality, ideology, and political knowledge scores.

B.7 Implications of GPT-3 Formatting Errors

In a small minority of cases, GPT-3 autocompletes sentences before summarizing them. Therefore, a respondent writing “gun control” may receive a Likert item stating “is importantI believe gun control is an important issue.” This was relatively rare across the studies (7% in Experiment 1; 5% in Experiment 2; 15% in Experiment 3), and the gist of the survey question is still discernible in most cases. However, we still think it is worth considering whether (1) these errors are distributed evenly across conditions and (2) these errors condition treatment effects. Given that our API calls are occurring before treatment assignment, we ought to expect minimal differences across conditions. This is exactly what we find. Errors are balanced across conditions, with one marginally significant exception. In the vast majority of cases, differences in errors are small and indistinguishable from zero.

Moving on to conditional effects, we find marginally significant evidence of greater attitude moderation among those in the directional and mixed information condition if they receive a properly formatted GPT-3 Likert summary in Experiment 1 ($\hat{\beta}$ = -.455; SE = .245; p < .10). In Experiment 2, we find marginally significant evidence that the “Con” effect on attitude certainty for the strong attitude is muted in the accuracy condition if participants receive a Likert item without this completion error ($\hat{\beta}$ = .590; SE = .352; p < .10). Finally, in Experiment 3, we find that receiving a “correct Likert item” decreases the effect of counter-attitudinal information on certainty ($\hat{\beta}$ = -.789; SE = .201; p < .001). The average treatment effect is negative and statistically significant among those with properly formatted Likert items ($\hat{\beta}$ = -.14; SE = .07; p = .05).

Taken together, the evidence suggests minimal effects of formatting errors on our findings. However, future iterations of this design should be mindful of creating prompts that minimize these errors. Encouraging longer open-ended responses that have a similar format might be one solution (i.e., “I believe that X”). Relative to the CloudResearch studies, the median number of characters in the Lucid open-ended question was 75 whereas the median number of characters in Experiment 1 and 2 was 100 and 93, respectively. In the Lucid study, approximately 36% of responses began with “I think” or “I believe,” whereas this number was 68% and 67% in Experiment 1 and 2, respectively. Future studies could encourage longer and more consistently formatted open-ended responses to produce lower error rates. Fine-tuning GPT-3 can also help reduce error rates. Using correctly formatted data from Experiments 1-3 as training data, we have found that error rates substantially decrease.

TABLE B3 . Balance in GPT-3 Errors

	Experiment 1	Experiment 2	Experiment 3
(Intercept)	0.933***	0.955***	0.852***
	(0.013)	(0.009)	(0.012)
Con	0.003	-0.008	
	(0.015)	(0.011)	
Mixed	0.010	-0.016	
	(0.015)	(0.011)	
Directional	-0.021+	0.005	
	(0.012)	(0.009)	
Counter-Attitudinal Argument			−0.025
			(0.019)
N	1782	2362	1895
R^2	0.002	0.001	0.0002

Note: Statistical significance levels: + p<.10; * p< .05; ** p<.01; *** p<.001

TABLE B4 . Conditional Effects of GPT-3 Errors (Experiment 1)

	Attitude Strength (1-7)	Attitude Certainty (0-100)
(Intercept)	6.924***	97.317***
	(0.076)	(0.706)
Pre-Treatment Attitudes	0.270***	0.039
	(0.063)	(0.191)
Pre-Treatment Certainty	0.181***	6.133***
	(0.034)	(0.473)
Correct Likert	-0.086	-0.217
	(0.081)	(0.733)
Con	-0.056	-0.217
	(0.079)	(0.884)
Mixed	-0.268	-0.542
	(0.198)	(0.880)
Directional	-0.190	-0.525
	(0.168)	(0.835)
Correct Likert × Con	0.045	0.209
	(0.089)	(0.941)
Correct Likert × Mixed	0.264	0.579
	(0.199)	(0.910)
Correct Likert × Directional	0.220	0.627
	(0.171)	(0.868)
Con × Directional	0.106	0.660
	(0.196)	(1.832)
Mixed × Directional	0.419+	0.666
	(0.243)	(1.073)
Correct Likert × Con × Directional	-0.123	-0.857
	(0.204)	(1.926)
Correct Likert × Mixed × Directional	-0.455+	-1.687
	(0.245)	(1.179)
N	1768	1772
R^2	0.440	0.719

Note: Statistical significance levels: + p<.10; * p< .05; ** p<.01; *** p<.001

TABLE B5 . Conditional Effects of GPT-3 Errors (Experiment 2)

	Attitude Strength (1-7)	Attitude Certainty (1-9)
(Intercept)	5.796***	7.473***
	(0.154)	(0.183)
Con	0.248	−0.650+
	(0.205)	(0.344)
Mixed	0.282	−0.244
	(0.192)	(0.286)
Correct Likert	0.101	−0.030
	(0.162)	(0.173)
Directional Prime	−0.076	−0.232
	(0.265)	(0.261)
Pre-Treatment Strength	0.480***	0.230**
	(0.123)	(0.083)
Pre-Treatment Certainty	0.550***	1.173***
	(0.116)	(0.126)
Con × Correct Likert	−0.185	0.590+
	(0.231)	(0.352)
Mixed × Correct Likert	−0.228	0.101
	(0.213)	(0.295)
Con × Directional Prime	0.107	0.594
	(0.309)	(0.436)
Mixed × Directional Prime	−0.025	0.263
	(0.311)	(0.398)
Correct Likert × Directional Prime	0.198	0.289
	(0.280)	(0.268)
Con × Correct Likert × Directional Prime	−0.275	−0.729
	(0.336)	(0.449)
Mixed × Correct Likert × Directional Prime	−0.070	−0.230
	(0.333)	(0.411)
N	1122	1126
R^2	0.214	0.348

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Note: Statistical significance levels: + p<.10; * p< .05; ** p<.01; *** p<.001

TABLE B6 . Conditional Effects of GPT-3 Errors (Experiment 3)

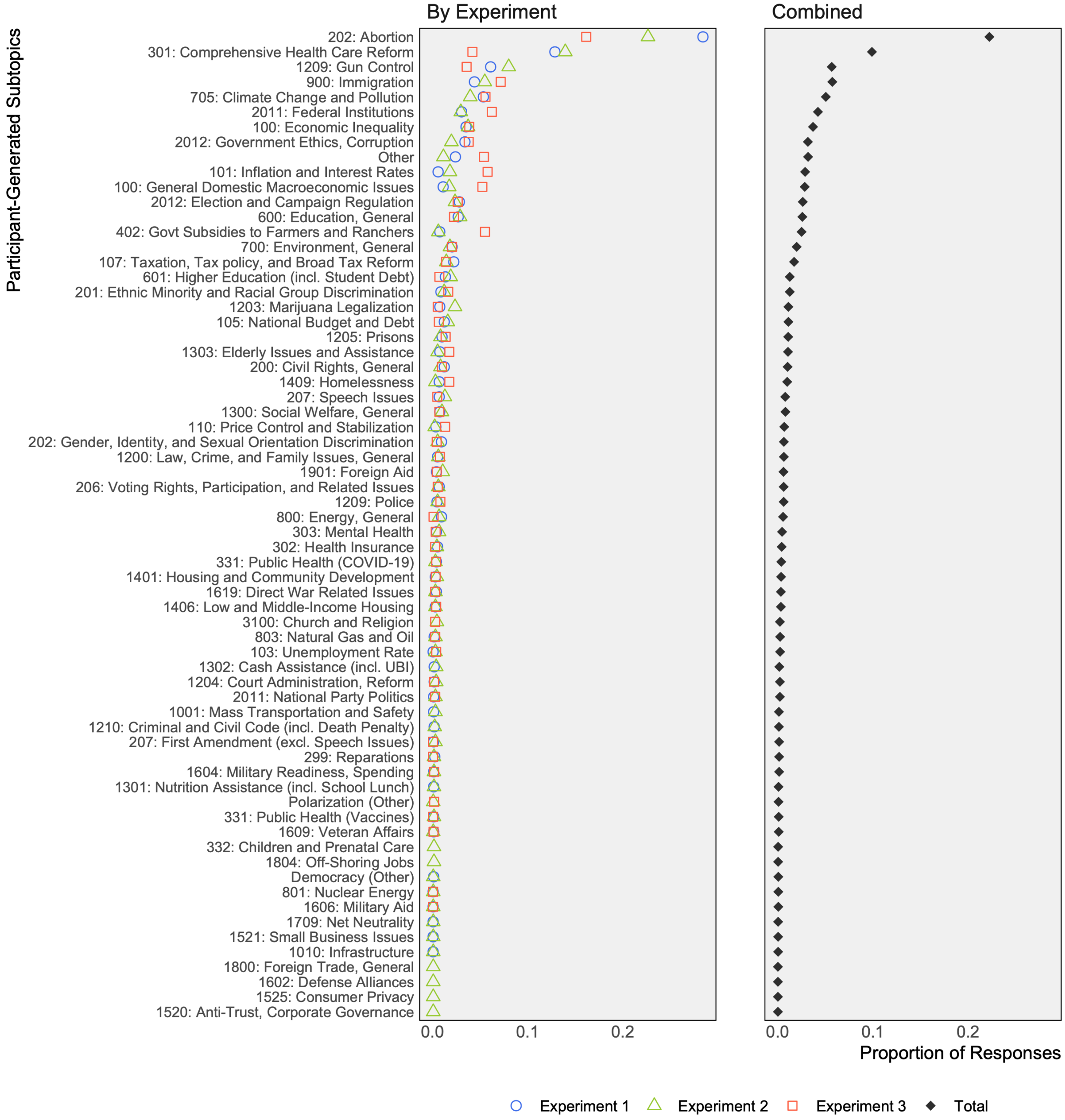
	Attitude Strength (1-7)	Attitude Certainty (1-9)
(Intercept)	5.255***	6.752***
	(0.138)	(0.150)
Treatment	−0.596**	0.643***
	(0.216)	(0.187)
Correct Likert	0.969***	1.087***
	(0.151)	(0.159)
Age	0.117*	0.315***
	(0.048)	(0.037)
Education	0.073	0.037
	(0.049)	(0.039)
Income	0.041	0.116**
	(0.046)	(0.036)
Ideology	−0.054	−0.062
	(0.058)	(0.049)
White	0.115*	0.092*
	(0.047)	(0.039)
Partisanship	−0.097+	−0.018
	(0.056)	(0.047)
Treatment × Correct Likert	0.147	−0.789***
	(0.236)	(0.201)
N	1848	1852
R2	0.080	0.105

Note: Statistical significance levels: + p<.10; * p< .05; ** p<.01; *** p<.001

B.8 Distribution of Issue Subtopics Across Experiments

Figure B7 illustrates the diversity of topics reflected in participants' self-reports of their core issue positions. Ryan and Ehlinger (2019, chapter 4) have recently explored the potential of open-ended survey questions aimed at eliciting important issue opinions for uncovering *issue publics* — small constituencies who are heavily invested in and who possess sophisticated opinions about niche political topics. Employing the Comparative Agendas Project (CAP) (2019) coding scheme, we find that our participants' responses across all experiments map onto 57 unique CAP subtopics. A handful of topics are further disaggregated (i.e., "202: Abortion" and "202: Gender, Identity, and Sexual Orientation Discrimination"), resulting in the 66 subtopics displayed in Figure B7 .

FIGURE B7 . Frequency of Participant-Generated Subtopics



On the one hand, the issues that top our list should come as no surprise. The five most common topics — abortion, health care reform, gun control, immigration, and climate change and pollution — collectively account for just under 50% of the responses across all experiments and are highly salient issues nationwide. On the other hand, no topic accounts for more than a quarter of responses across all experiments, and the prevalence of each topic varied markedly between experiments. We cannot know whether a respondent who thinks of abortion when inquired about the political issues most important to them holds equally crystallized and accessible views about gun control and immigration, let alone climate change, education, or a less conventional topic. Nor should we presume that such a respondent would be equally likely to react affectively to persuasive messages regardless which of these topics were presented in the study. Those respondents who wrote about a plethora of idiosyncratic issue areas, ranging from veteran affairs to police reform, may possess deeply held and passionate attitudes on these matters rooted in personal life experiences, yet may feel lukewarm about topics more frequently or more recently invoked in national political discourse. By eliciting open-ended responses, we avoid making the assumption that all or even most of our participants innately possess important attitudes on any one topic.

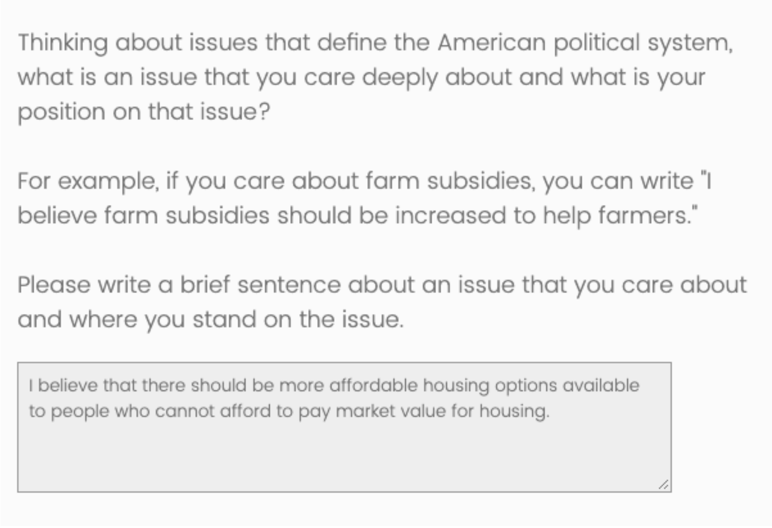
C Experimental Materials

C.1 Question Wording

- Attitude strength (1-7): To what extent do you agree or disagree with the following statement? [GPT-3 summary] (Strongly disagree, disagree, somewhat disagree, neither agree nor disagree, somewhat agree, agree, strongly agree)
- Attitude certainty (0-100): How certain are you regarding this position?[GPT-3 summary] When giving a score, keep in mind: - 100 = Complete certainty. You could not be more sure of where you stand. - 75 ≈ Very sure about where you stand. - 50 ≈ Pretty sure about where you stand. - 25 ≈ Mildly sure about where you stand. - 0 = Complete uncertainty. You have a belief, but you are very unsure if you should hold it.
- Attitude certainty scale (1-9): Each item is measured using a 9-point scale ranging from 1 (Not at all certain) to 9 (Very certain). Only endpoints of the scale are labeled.
  - How certain are you that your attitude on this issue is the correct attitude to have?
  - How certain are you that other people should have the same attitude as you on this issue?
  - How certain are you that of all the possible attitudes one might have on this issue, your attitude reflects the right way to think and feel about the issue?
  - How certain are you that you know what your true attitude on this issue really is?
  - How certain are you that the attitude you expressed on this issue really reflects your true thoughts and feelings?
  - To what extent is your true attitude on this issue clear in your mind?
  - How certain are you that the attitude you just expressed on this issue is really the attitude you have?
- Attitude Duration: How long have you held this belief with this level of certainty?
  - Since beginning this survey
  - In the last week
  - In the last month
  - In the last 2 to 6 months
  - In the last 7 to 12 months
  - In the last 2 to 3 years
  - Longer than 4 years
- Discussion Frequency: In the past 6 months, how often have you debated with family members, peers, or people online about this issue?
  - More than once a week
  - Once a week
  - Once or twice a week
  - Once a month
  - A few times
  - Have not debated this issue in past 6 months

C.2 Survey Interface (Experiments 1 and 2)

FIGURE C1 . Requesting Strong Attitudes (Experiments 1, 2, and 3)



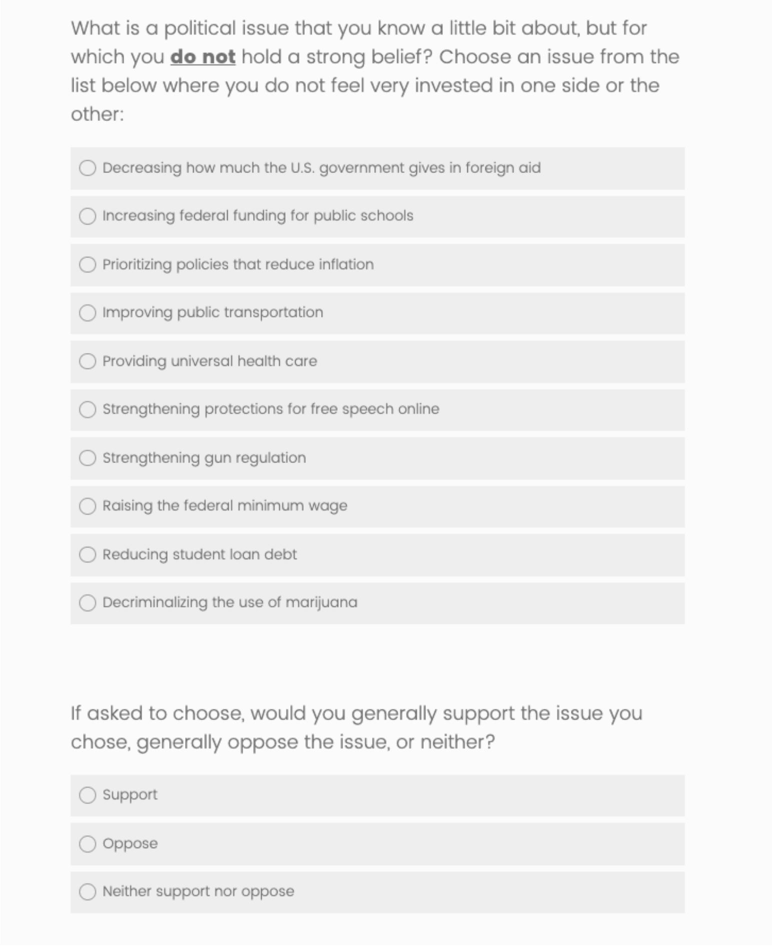
Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue?

For example, if you care about farm subsidies, you can write "I believe farm subsidies should be increased to help farmers."

Please write a brief sentence about an issue that you care about and where you stand on the issue.

I believe that there should be more affordable housing options available to people who cannot afford to pay market value for housing.

FIGURE C2 . Requesting Weak Attitudes (Experiment 2)



What is a political issue that you know a little bit about, but for which you **do not** hold a strong belief? Choose an issue from the list below where you do not feel very invested in one side or the other:

○ Decreasing how much the U.S. government gives in foreign aid

○ Increasing federal funding for public schools

○ Prioritizing policies that reduce inflation

○ Improving public transportation

○ Providing universal health care

○ Strengthening protections for free speech online

○ Strengthening gun regulation

○ Raising the federal minimum wage

○ Reducing student loan debt

○ Decriminalizing the use of marijuana

If asked to choose, would you generally support the issue you chose, generally oppose the issue, or neither?

○ Support

○ Oppose

○ Neither support nor oppose

Figure C1 displays the survey interface for requesting open-ended responses of respondents' deeply-held attitudes, used in Experiments 1, 2, and 3). The response used concerning affordable housing is a real respondents' input from Experiment 2. Figure C2 displays the survey interface asking respondents to select a topic about which they possess a less important attitude.

In Experiments 1 and 2, respondents were randomly assigned to view one of two sets of instructions prior to the thought listing task, one written to prime accuracy motivations and the other written to prime directional motivations. The accuracy prime text was as follows:

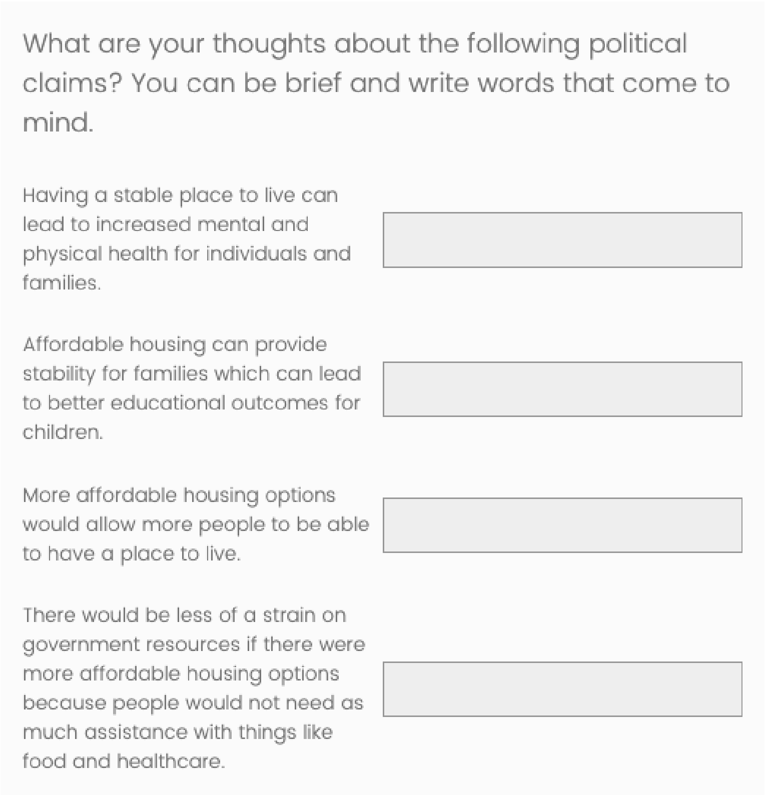
> On the next page, you will be asked to read a set of statements. As you carefully read each statement, assess how accurate its claims are to the best of your ability.
>
> It is highly important that you **ignore** any personal feelings or emotions you might experience in response to reading these statements. Focus only on determining the truth of each statement.
>
> Here are some questions you might think about:
>
> - Does this claim make logical sense?
>
> - Does this statement seem to reflect reality?
>
> - Can I think of strong supporting evidence for this claim?
>
> - Can I think of strong counterarguments that invalidate this claim?
>
> Then, use the space provided to write a few words or even a sentence explaining your thoughts.

The directional prime text was as follows:

> On the next page, you will be asked to read a set of statements. As you read each statement, think about whether or not its claims align with your personal and political beliefs.
>
> **Do not** worry about determining if each statement is accurate. What matters is whether you are able to maintain consistency with your own worldview and the beliefs you've developed over the course of your life. Focus on understanding what each statement means to you given your personal and political beliefs.
>
> Here are some questions you might think about:
>
> - Does this claim resonate with my beliefs on this issue?
>
> - Does this claim resonate with my moral and social values?
>
> - Does this claim conflict with my political affiliations and commitments?
>
> - How would people who share my political affiliations and commitments feel about this claim?
>
> Then, use the space provided to write a few words or even a sentence explaining your thoughts.

Figure C3 displays the survey interface for the thought listing task. The example arguments shown are pro arguments generated by GPT-3 in response to the affordable housing prompt shown in the earlier figure, "Requesting Strong Attitudes."

FIGURE C3 . Thought Listing Task (Experiments 1 and 2)



What are your thoughts about the following political claims? You can be brief and write words that come to mind.

Having a stable place to live can lead to increased mental and physical health for individuals and families.

Affordable housing can provide stability for families which can lead to better educational outcomes for children.

More affordable housing options would allow more people to be able to have a place to live.

There would be less of a strain on government resources if there were more affordable housing options because people would not need as much assistance with things like food and healthcare.

C.3 GPT-3 Details

We provide the details of our various API calls here. The temperature parameter varies the extent to which the output is deterministic versus stochastic. Lower scores tend to generate similar output across different API calls. Max tokens captures the maximum amount of text tokens produced by GPT-3. The frequency penalty is a parameter that limits the tendency for GPT-3 to repeat output, whereas the presence penalty is a parameter that varies whether the model should produce more "novel" text predictions. OpenAI has released several models that vary in complexity and cost. In all of our experiments, we choose the "davinci" class of GPT-3 models, which is the most powerful class of models available.

Likert API call:

- Model: Davinci-002
- Prompt: "Summarize this person's belief in the first person in one sentence (always begin with the word I believe and stick with what the person says):"
- Temperature: 0.7
- Max Tokens: 50
- Frequency Penalty: 0
- Presence Penalty: 0

Pros and Cons API call:

- Model: Davinci-002
- Prompt: "Write four pros and four cons for the following argument in a JSON format. Each pro and con should be labeled pro1, pro2, pro3, pro4, con1, con2, con3, con4."
- Temperature: 0.7
- Max Tokens: 250
- Frequency Penalty: 1
- Presence Penalty: 1

Counterattitudinal Argument API call:

- Model: Davinci-003
- Prompt: "Write a paragraph-long passionate rebuttal of the following argument. The author should strongly disagree with the statement."
- Temperature: 0
- Max Tokens: 256
- Frequency Penalty: 0
- Presence Penalty: 0

Open-Ended Response Quality Check API call:

- Model: Davinci-003
- Prompt: "Is the following response relevant to the prompt? Always generate a number (0 = low quality, .5 = medium quality, 1 = high quality).

Examples: Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response: I believe in abortion. {quality: ".5"}
Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response: I believe if California stops handing out free money, they will go broke {quality: ".5"}
Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response: I believe foods free help a lot of people {quality: "0"}
Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response: Inflation has caused me to lose my house and car {quality: ".5"}
Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response: Itś a issue about money {quality: "0"}
Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response: I believe that healthcare is a human right and that everyone deserves access to quality care. {quality: "1"}
Prompt: Thinking about issues that define the American political system, what is an issue that you care deeply about and what is your position on that issue? Response:"

- Temperature: 0
- Max Tokens: 50
- Frequency Penalty: 0
- Presence Penalty: 0

C.4 Sample Composition

TABLE C1 . Sample Composition

	Experiment 1	Experiment 2	Experiment 3
Gender (F)	0.58	0.54	0.52
Age	43.11	43.40	45.02
Income	5.37	5.66	8.31
Education	4.30	4.38	4.39
Ideology (7-point scale)	2.68	2.79	3.18
Democrat	0.46	0.48	0.34
Republican	0.21	0.24	0.25
Independent	0.33	0.29	0.41
White	0.80	0.80	0.63
Latino	0.06	0.06	0.16
Black	0.06	0.07	0.12
AAPI	0.07	0.07	0.05

Note: Mean estimates for demographic categories. Racial and ethnic demographic estimates for Experiment 1 and 2 were obtained by recontacting a random sample of participants from each study (N = 517).

C.5 Example of GPT-3 Output (Experiment 3)

Though the full set of arguments can be found here (bit.ly/3XjLjZC), we provide three examples of the arguments that GPT-3 produced in Experiment 3. As we describe in the manuscript, we instructed GPT-3 to write longer and more affectively charged arguments. For example, an individual who wrote "I care about gun control. I do not understand why so many people, especially that have a criminal past or younger children have access to them." received the following response:

> Gun control is an important issue, but it is not the only factor in preventing gun violence. It is not reasonable to assume that everyone with a criminal past or young children should be denied access to guns. People with criminal records can still be responsible gun owners, and young children can be taught to use guns safely and responsibly. Furthermore, denying access to guns does not guarantee that gun violence will be prevented. We need to focus on addressing the root causes of gun violence, such as poverty, mental health, and access to education, in order to create lasting change.

A participant who wrote "I care strongly about the lives of unborn children. Abortion SHOULD NOT be allowed unless there is a health issue with the mother or the possible the baby. The woman was raped, or incest." received the following counter-argument:

> I vehemently disagree with the notion that abortion should only be allowed in cases of health issues with the mother or the baby, or if the woman was raped or experienced incest. This is an incredibly narrow view of the issue that fails to take into account the myriad of other circumstances that can lead a woman to seek an abortion. Women should have the right to make decisions about their own bodies and lives, and should not be forced to carry a pregnancy to term against their will. Furthermore, it is not the place of the government to dictate what a woman can and cannot do with her body. Everyone should have the right to make decisions about their own lives and bodies, and that includes the right to choose whether or not to have an abortion.

A participant who wrote "Health Care should be a priority. No one in USA should have to worry about it. We should have Universal healthcare, period‼" received this argument:

> While I agree that health care should be a priority, I strongly disagree with the idea that universal healthcare should be the only option. Universal healthcare is an expensive and inefficient system that would place an undue burden on taxpayers. Furthermore, it would limit the choices of individuals and families when it comes to their health care needs. Instead, we should focus on providing more affordable and accessible health care options that allow individuals and families to make their own decisions about their health care. This would ensure that everyone has access to quality health care without sacrificing their freedom of choice.